이학박사학위논문

A Statistical Analysis for Next-Generation Sequencing Data with a Small Number of Samples

자료수가 적은 차세대 염기서열자료의 통계적 분석

2014년 2월

서울대학교 대학원

(협동)생물정보학과

김 정 수

A Statistical Analysis for Next-Generation Sequencing Data with a Small Number of Samples

지도교수 박태성

이 논문을 이학박사 학위논문으로 제출함

2014 년 2 월

서울대학교 대학원

(협동)생물정보학과

김 정 수

김정수의 이학박사 학위논문을 인준함

2014 년 2 월

위 원 장 김 선 (인)

부위원장 박 태 성 (인)

위 원 천 종 식 (인)

위 원 이 승 연 (인)

위 원 김 희 발 (인)

학위논문 원문제공 서비스에 대한 동의서

본인의 학위논문에 대하여 서울대학교가 아래와 같이 학위논문 저작물을 제공하는 것에 동의합니다.

1. 동의사항

①본인의 논문을 보존이나 인터넷 등을 통한 온라인 서비스 목적으로 복제할 경우 저작물의 내용을 변경하지 않는 범위 내에서의 복제를 허용합니다.

②본인의 논문을 디지털화하여 인터넷 등 정보통신망을 통한 논문의 일부 또는 전부의 복제.배포 및 전송 시 무료로 제공하는 것에 동의합니다.

2. 개인(저작자)의 의무

본 논문의 저작권을 타인에게 양도하거나 또는 출판을 허락하는 등 동의 내용을 변경하고자 할 때는 소속대학(원)에 공개의 유보 또는 해지를 즉시 통보하겠습니다.

3. 서울대학교의 의무

①서울대학교는 본 논문을 외부에 제공할 경우 저작권 보호장치(DRM)를 사용하여야 합니다.

②서울대학교는 본 논문에 대한 공개의 유보나 해지 신청 시 즉시 처리해야 합니다.

논문제목: A statistical analysis for next-generation sequencing data with a small number of samples (자료수가 적은 차세대염기서열 자료의 통계적 분석)

학위구분 : **석사 □.박사 ☑**
학 과 : (협동)생물정보학과
학 번 : 2006-30792
연 락 처 :010-5875-0906 (iedenkim@gmail.com)
저 작 자 : 김정수 (인)

제 출 일 : 2014 년 2월 3일

서울대학교총장 귀하

A Statistical Analysis for Next-Generation Sequencing Data with a Small Number of Samples

by

Jungsoo Gim

A thesis
submitted in fulfillment of the requirement
for the degree of Doctor of Philosophy
in
Bioinformatics

Interdisciplinary Program for Bioinformatics

College of Natural Sciences

Seoul National University

Feb, 2014

ABSTRACT

A Statistical Analysis for Next-Generation Sequencing Data with a Small Number of Samples

Jungsoo Gim

Interdisciplinary Program for Bioinformatics

The Graduate School

Seoul National University

With an advance of technology, new methods to meet a more suitable analysis that ever has been made, need to be developed. Since the microarray technology had been developed, plenty of methods have been invented, from genome-wide association analysis, which detects causative variants associated with diseases, to differential expression analysis, which identifies genes with dissimilar in abundance. In the early era, when the data was generated at great expense, researcher devoted to develop a method for the analysis of studies with small sample size. However, fast stabilization and incompleteness of the microarray technology lead many studies with larger sample size.

The efforts made by numerous scientists were concentrated on incorporating revisions into new methods for an analysis of microarray data. Therefore,

microarray technology has experienced fast stabilization. In microarray technology, the information of interest should be pre-acquired and placed on a limited space as a set of probes. Because of this property of microarray technology, there has been limits to the amount and the variety of information we can access. Thus it is more suitable for detecting common information rather than individual-specific information with microarray. Thus, rather than small sample studies, microarray technology dedicated to large sample studies to elucidate common phenomena observed in a large sample.

Next-generation sequencing (NGS) technology is inherently suitable for detecting individual information. It was a well match between NGS technology and the 'personalized' concept from the start of Human Genome Project. However, it is not easy to clarify the meaningful information from an individual data with a large amount of 1 base-pair resolution scale. Furthermore, relatively high cost and limited specimen availability often lead to studies with small samples (replicates). Eventually, to obtain results with significance from data with a small number of samples attracts researcher's attention.

In this thesis, the approaches to genomic data and transcriptomic data both with small sample sizes will be provided. Specifically, for genomic data analysis, a new strategy called multiphasic analysis is suggested. Applying the strategy to a Mendelian disease, the strategy shows how it efficiently weed out a disease-causing variant from various candidates.

For transcriptomic data analysis, a new method is proposed for analysis of

differential expression analyses between two classes, which can be applicable to RNA-Seq data with a small (even with non-replicated) number of replicates. the validity of the proposed method is provided by applying it to various real and simulated datasets and comparing the results to those obtained from other competing methods.

Keywords: NGS, RNA-Seq, Statistical analysis, Exome-Seq

Student number: 2006-30792

CONTENTS

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

This chapter gives a brief view of biological contents of interest in this thesis and a summary of revolutionary methodologies of last decade and their impact on biological thinking. The purpose and a scope of the thesis will be also stipulated in last section.

1.1 Background of Omics

The advent of the Human Genome Project captured the imagination of both scientists and the general public. Of course, however, it is almost impossible to fully summarize the human genome, in the sense that new discoveries will continue for many years, and to elucidate all the phenomena in question. The analysis of the genome itself will limit to extract all the meaningful information. Therefore, investigators needed to perform laboratory experiments in other layers of biology, such as transcripts and proteins, to glean more information than is apparent from genome analysis. For example, alternative splicing makes a significant degree of complexity in defining the proteome but one cannot detect it only with genome information.

There is a movement afoot in biology (Jain and Jain 2001). Some called it

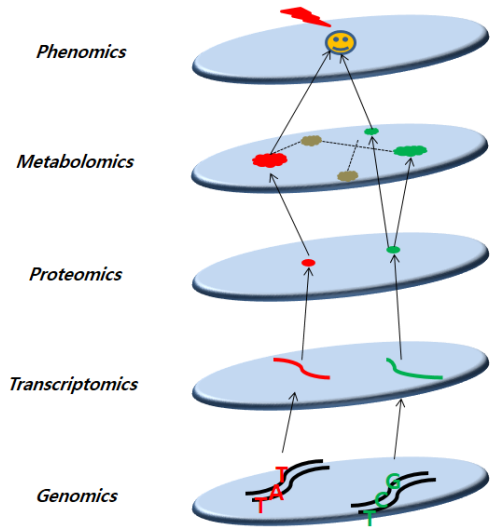


Figure 1-1 A schematic diagram of multi-level omics data.

integrative biology, but it has several synonyms. It is basically an approach to the studies of complex interaction of all layers (or levels) of biological information (Figure 1-1). To developing a model which will both describe the nature of the system and its systems properties, the most urgent priority is to analyze the data in each level. The basic idea in this philosophy is that each omics data is equally important and is worth studying each. For example, if one is to understand how a car functions, biologist would have studied the individual parts in isolation – the transmission; the ignition; the brakes, etc. The integrative or systems approach defines all of the elements in a system and then studies how each behaves in relation to the others as the system is functioning. Ultimately, this approach requires a more

accurate model in each level, which will eventually help the analysis in systems level.

1.1.1 Genomics

Genomics is an unusual scientific term because it definition varies from person to person. The root word *genome* is universally defined as the total DNA content of a haploid cell or half the DNA content of a diploid cell. You would think the discipline of genomics would be the study of genomes, but this simple definition is too simplistic. In one sense, all of biology is related to the study of genomes because an organism is shaped by its genome. However, most biologists would agree that disciplines such as anatomy and zoology should not be lumped into the current usage of genomics.

Genomics involves large data sets (about 3 billion base pairs for the human genome) and high-throughput methods (fast methods for collecting the data). Genomics includes sequencing DNA and collecting genome variations within a

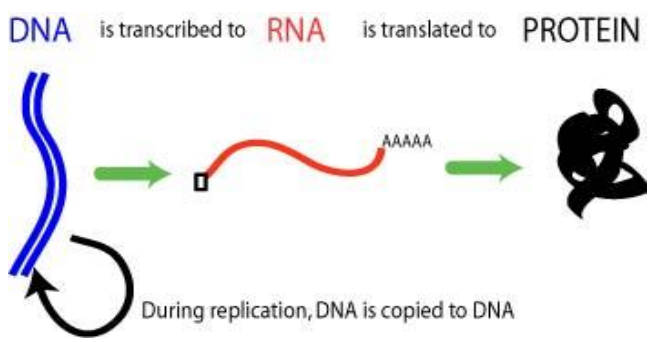


Figure 1-2 The Central Dogma.

(http://cnx.org/content/m11415/latest/)

population as well as transcriptional control of genes, since it is the originator fo the central dogma (Figure 1-2). Sometimes a broad definition of genomics, from DNA sequence analysis to an organism′s response to environmental perturbations, is used including other layers of omics. However, throughout this thesis, genomics is defined as a study of genome, informative sequences in DNA.

1.1.2 Transcriptomics

Genomic information written in DNA is literally an information, i.e., it cannot function to cell by itself but largely affects to an intermediate step called transcription. Transcription is placed in the early process of the central dogma (Figure 1-2). Since no observation has been made about direct translation from DNA to protein, transcriptome is laid a huge responsibility on gene function process.

Trasncriptomics is a study of the complete set of RNAs (transcriptomes) encoded by the genome of a specific cell or organism at a specific time or under a specific set of conditions. In transcriptome analyses, many things can be asked: gene expression, disease classification, functional annotation, and etc. For instance, gene expression analysis focuses on comparing gene expression patterns in tissues in which the particular treatment or condition is introduced.

Unlike the genome, the transcriptome has a dynamic property. Most of cells in a species contain the same genome regardless of the type of cell, stage of development or environmental conditions. Conversely, the transcriptome varies remarkably in differing circumstances due to different patterns of gene expression.

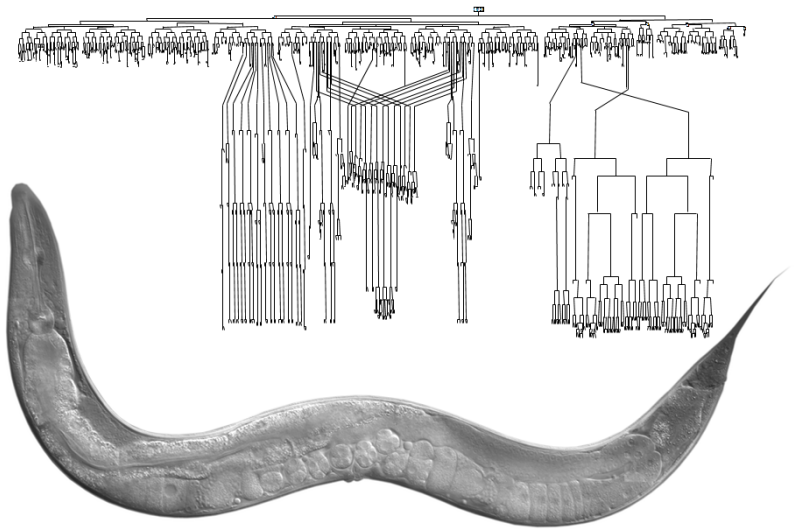


Figure 1-3 C. Elegans lineage.

Those difference results in different responses. As shown in Figure 1-3, genes can be classified into distinct cell types according to their different gene expression patterns, which affects different cell morphology (Figure 1-3).

1.1.3 Proteomics

Proteomics is the large-scale study of proteins, particularly their structures and functions. As will be described a bit detail in section 1.2.1, difficulty of microarray technology on protein had made researchers to investigate more on protein structures till so-called ′post-genome era′. Recently, however, functions of proteins are getting attention and the next big challenge is to understand proteomes.

Proteomics is a very difficult field because we lack methods for completely defining proteomes, partly because they present a moving target. In addition to cell-to-cell variation, any given cell will change its proteome over time. Nevertheless many struggles have achieved protein identification and quantification, and also their function to cells using newer methods that utilize tandem mass spectroscopy.

As tools become more precise, studying new properties and protein functions have been conducted. The term 'function' was used to describe what a protein dose, but a growing number of scientists fell that function is too limited. A consortium of genomic groups called 'Gene Ontology' have collaborated to break the term function into three more specific ones (Ashburner, Ball et al. 2000):

1. Biological process (why − why is this being done? e.g., movement of cell)
2. Molecular function (what − what kind of molecule is this? e.g., ATPase)
3. Cellular component (where − where is this located? e.g., ribosome)

The area of proteomics is like the front line of an avalanche where advances protrude from different places along the edge but each advancing edge is followed very quickly by the rest of a fast-paced progression.

1.2 Technologies Measuring Omics

It has been remarkably successful to elucidate the phenomena in cells and tissues with microarray-based technology. However there are a number of shortcomings to this technology. Both sensitivity and specificity can be low with microarrays. This section briefly provides a detail of each technology.

1.2.1 Microarray technology

A micoarray is a multiplex lab-on-a-chip, a solid substrate assaying large amounts of biological materials using high-throughput screening miniaturized, multiplexed and parallel processing and detection methods. According to the biological materials placed on a chip, the microarray is called in distinct names: cDNA microarrays, oligonucleotide microarrays, SNP microarrays, protein microarrays, and etc.

The detailed descriptions on microarray is beyond the thesis but the key point that should be pointed out is its dependency nature. The technology solely depends on the information on an array. For the information not put on an array, there is no way out to find these ′additional information′ whatsoever (Figure 1-4).

Additionally, microarrays restrict the expression profiling data to specific annotations and contents with negatively affected by the low dynamic range of existing microarray platforms. Those difficulties are resolved in next-generation sequencing technology at the price of high cost.

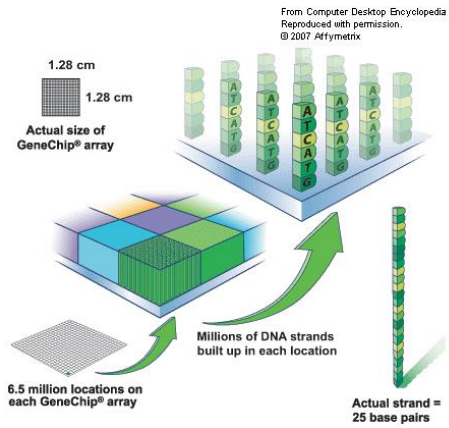


Figure 1-4 An oligomer microarray of GeneChip

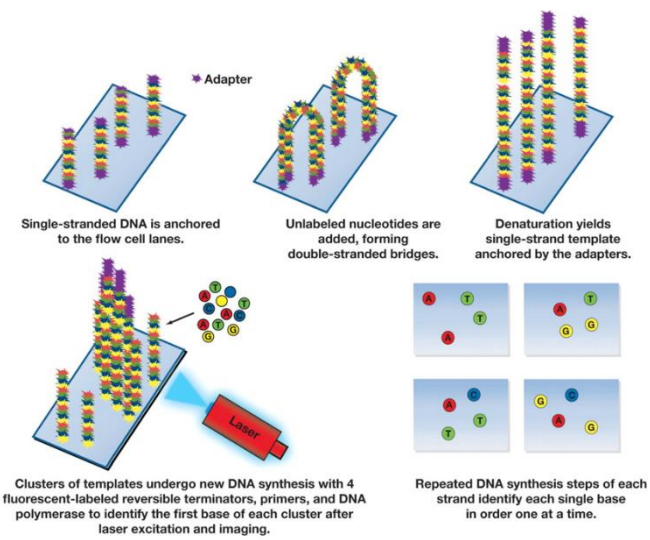


Figure 1-5 NGS technology of Illumina/Solexa

1.2.2 Next-generation sequencing (NGS) technology

There had been strong demand for low-cost and high throughput sequencing, producing thousands or millions of sequences simultaneously. Since the first of the NGS was developed in the 1990s, many methods have been developed in various researchers (Shaffer 2007). All these high-throughput NGS technologies are intended to lower the cost of DNA sequencing beyond what is possible with standard dye-terminator methods, such as Sanger sequencing. Despite of many effort, however, it still is far expensive than microarray technology. Table 1-1 shows characteristics of each technology.

Table 1-1. Notable features of NGS and microarray technologies

	NGS	**Microarray**
System type	Open system	Closed system
Depth of sample coverage	Higher	Lower
Appropriate throughput	Getting higher	higher
Cost efficiency with multiple samples	More expensive per sample	Less expensive per sample
Ease of data handling/analysis	Complex annotation and sorting of massive sequence read	Simple analysis of signal intensities
Information availability	Based on sample	Based on probes on an array
Recommended application	In-depth studies of unknown private diversity	Routine studies of functional gene diversity across many samples

1.2.3 Mass Spectrometry (MS)

There exists several methods measuring proteins, such as two-dimensional (2D) gel electrophoresis and mass spectrometry (MS). MS is a preferred analytical technique that produces spectra of the masses of the atoms or molecular comprising a sample of interest. The spectra are used to determine the elemental or isotopic signature of a sample, the masses of particles and of molecules, and to elucidate the chemical structures of molecules, such as peptides of proteins.

There are two key components to MS. The first is that all proteins can be sorted based on a mass to charge ratio (i.e., the molecular weight divided by the charge on

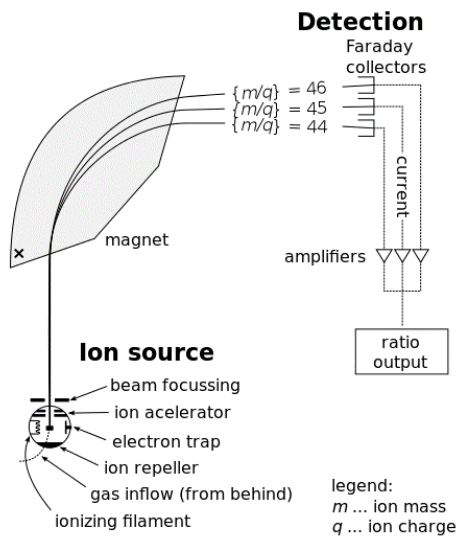


Figure 1-6 A schematic diagram of mass spectrometry. MS consists of several steps: ionization, separation, activation and mass determination step.

this protein). The second is that proteins can be broken into peptide fragments, facilitating the identification of each protein.

1.3 Analysis of NGS Data with Small Sample Size

This section briefly summarizes the characteristics of NGS data and describes the necessity for the analysis with a small sample.

1.3.1 Necessity for the small sample analysis

We are frequently left with small samples due to several reasons in many cases. One might give a clue to the reasons with a sequencing cost. Even though the development goes towards large sample size (as microarray did), and barcoding and multiplexing create opportunities to analyze more samples at a fixed cost, as of today NGS experiments are often too expensive to allow extensive replication or samples.

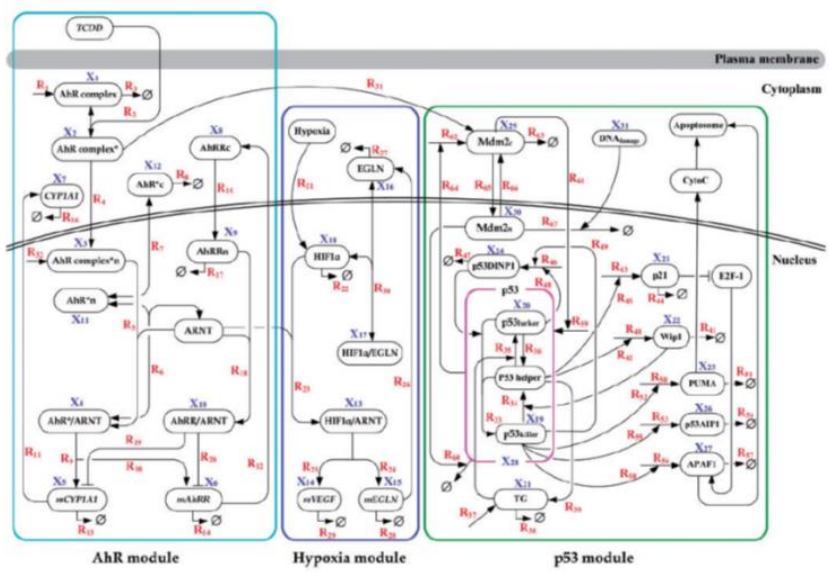


Figure 1-7 A schematic diagram of signaling crosstalks. Three different signaling pathways are interacting each other. Each individual might react differently to intakes of TCDD.

Limited specimen availability also limits researcher to studies with small sample. As pointed out in earlier sections, microarray technology needs pre-acquired information to obtain information of the cells based on the pre-acquired information, while NGS technology needs those information to distinguish cells' properties from the information. If you take careful considerations on this, you might come to a conclusion that microarray is suitable for the analysis of common exploratory factors associated to a specific phenomenon. However, NGS observes private factors owned by individual-specific samples, each of which associated to a common phenomenon.

No doubt we usually come to a more accurate decision with more information. However, how to achieve to it is a different issue. Let's think of a system consists of three different signaling pathways: AhR pathway, hypoxia pathway, and p53 pathway (Figure 1-7). Intakes of TCDD in body induces AhR pathway, and thus down-stream signaling pathway is activated. As a result, CYP1A1 gene and AhRR gene is expressed to react to TCDD intake. This AhR pathway, induced by TCDD intake, cross-talks with other pathways, whose basal status might be different in each individual (Gim, Kim et al. 2010). For whom with high hypoxia basal status in normal, less transcriptional outcome would be observed because of hypoxia signaling crosstalk. With different basal level of hypoxia, different outcomes in AhR target genes expression is expected.

What if we perform differential expression analysis between TCDD intakes group and normal group with gene expression data with individual-specific status ignored? In such analysis, hypoxia status of each individual is likely to be neglected.

Without considering other factors, which significantly affect the TCDD down-stream signaling pathway, exact TCDD responses are hardly observed.

There lays a necessity for the small sample analysis. We need to consider exact characteristics of each sample, combine the results made from each sample, and conclude with each sample. In brief, regardless of sequencing cost drop, we can faced with the studies with small samples and we should clearly clarify the information of interest from the small sample analysis.

1.3.2 Purpose and novelty of this study

The main purpose of this thesis is to provide the methods analyzing NGS data with small sample size. To this end, the thesis focus on two studies. One is a study to prioritize the causative gene responsible to non-syndromic hearing loss (NSHL), a Mendelian disease. The other is a study of differential expression (DE) analysis of RNA-Seq data with a small number of replicates.

The first example investigated whether linkage analysis could be performed using genotypes inferred from whole-exome sequencing data, removing the need for the array-based genotyping step. The first study shows that the proposed strategy effectively reduces the search space of candidate variants for the disease.

In the second study, a new method is proposed for the analysis of RNA-Seq experiments when no replicates are available for each condition (or class). The most of existing methods are applicable to data with replicates. Thus differential expression analysis between two different conditions, each of which has no replicate,

is not feasible using existing methods. A proposed method is applied various simulation and real datasets.

1.4 Outline of the thesis

Following chapter will be devoted to cover genomic data analysis. Not whole area of genomics is covered in this thesis, but small samples of a family with a Mendelian disease measured with whole-exome sequencing will be discussed. In chapter 3, transcriptomic data analysis will follow with examples of two-class comparison with small number of replicates. A method called ′LPEseq′ is developed, which is designed for a differential expression analysis of RNA-Seq experiments with a limited number of replicates per condition. Conclusion and discussion will be provided in each chapter. Overall concluding remarks are presented in the last chapter of the thesis. The method is implemented in R language. Appendix provides source code of the *LPEseq* package implemented in R language and a practical manual for *LPEseq* package.

CHAPTER 2

GENOMIC DATA ANALYSIS

This chapter provides an overview of the search for genetic variants that influence the susceptibility of an individual to a disease, from the conventional family-based studies to the recent excitement of genome-wide association (GWA) studies. I then discusses the concepts behind the identification of common variants as disease causal factors and contrast them to the basic ideas that underlie individual-specific rare variant hypothesis. From the discussion, I concentrate on advantage of family-based studies with small samples. The following section is about a review of identification methods for detecting causal variants for family-based studies. A multiphasic approach is then introduced using whole exome sequencing data of a nonsyndromic hearing loss (NSHL) family. I emphasize that the strategy is not methodologically innovative, but it is innovative in aspect of data usability.

2.1 Introduction

Figure 2-1 A schematic example of Single Nucleotide Polymorphism (SNP)

Various genomic variants are associated with diseases prevalence. Unlike SNP-chip microarray, re-sequencing technologies can detect individual-specific genomic variants, such as copy number variation (CNV), single nucleotide variant (SNV), etc. Both population-based and family-based designs are commonly used in genetic association studies to identify causal genomic variants that underlie complex diseases. In this introduction section, I define what genomic variants are and describes which of two designs (population-based or family-based) suit the analysis of small samples.

2.1.1 Genomic variants and disease association

How much variation is there in the human genome, and are these variations informative to identifying disease mechanism? In the middle of the Human Genome Project is this question of variation in the human genome.

Ahead of microsatellite variations in a genome, a type of polymorphism has

gained more attention in recent decades, called single nucleotide polymorphisms, or (SNPs). SNPs are single bases at a particular locus where individuals have difference in their sequences. SNPs are another form of genomic variation in a population that may occur anywhere in the genome, and each person will have many SNPs.

There might be many ideas of how SNP data will be useful. One of the biggest interests is that SNPs are reliable markers that may allow us to determine which combination of coding alleles are associated with particular diseases. For this discussion, three confusing terminologies are described here – linkage, linkage disequilibrium, and haplotype. Linkage refers to how close two loci are to each other on a chromosome. If they are near each other, we say the two loci are linked. Linkage disequilibrium describes alleles rather than loci. If two alleles (or two SNPs) tend to inherited together more often than would be predicted, we say the alleles are in linkage disequilibrium (i.e., inherited together more often than other possible allele pairs). Haplotype refers to the set of alleles on one particular chromosome. Each person has two haplotypes in a given regions, and each haplotype will be passed on a complete unit unless recombination occurs to separate this particular set of alleles to form two new haplotypes.

Let's look at a hypothetical example where two SNPs and one gene are associated with a monogenic disease (Figure 2-2). Notice how the two different populations have different ratios of the three possible genotypes, which is what you would expect to see in a population that experiences a higher-than-average incidence

Frequency		Genomic DNA	SNP haplotype	Phenotype
P1	P2			
78%	52%		1 - 2 1 - 2	wt
15%	37%		1 - 2 1′ – 2′	wt
7%	11%		1′ – 2′ 1′ – 2′	disease

Figure 2-2 Comparison of SNP data for two populations. Two population (P1 and P2) have different frequencies of a single-locus disease. Included in this figure are diagrams of the genomic DNA showing two SNPs (1' and 2') and a recessive allele (denoted light grey) that are in linkage disequilibrium.

of the disease. Even if you did not know minor allele (denoted light grey in Figure 2-2) was the causative agent, you could evaluate the SNP haplotypes and deduce that SNPS 1' and 2' are linked to the disease allele, and that is a recessive disease since an individual has to be homozygous to contract it.

Another form of genomic variants are copy number variations (CNVs). CNVs are alterations of the DNA of a genome that results in the cell having an abnormal or, for certain genes, a normal variation in the number of copies of one or more sections of the DNA. For example, the chromosome that normally has sections in order as A-B-C-D might instead have sections A-B-C-C-D (a duplication of "C") or A-B-D (a deletion of "C").

Like other types of genomic variation, some CNVs are reported that they are

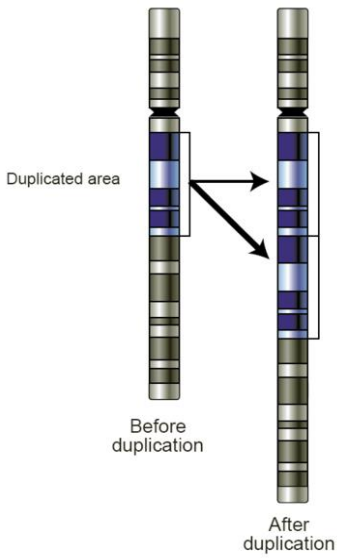


Figure 2-3 Gene duplication diagram. This gene duplication a copy-number variation. The chromosome now has two copies of this section of DNA.

(From National Human Genome Research Institute)

associated with susceptibility or resistance to diseases. Gene copy number can be elevated in cancer cells. For instance, the number of the gene called EGFR are significantly higher than normal in lung cancer.

2.1.2 Population-based vs. Family-based studies

Two fundamentally different designs are used in genetic association studies: family design and population design (Carrasquillo, McCallion et al. 2002, Saito and Kamatani 2002, Laird and Lange 2006, Ott, Kamatani et al. 2011). The population and family designs, which have different strengths and weaknesses, should be viewed as complementary and not as competitive in the effort to overcome the challenges of association studies for complex diseases.

In terms of statistical power, the differences between the two approaches are generally small when the use of trios in family designs is compared to case-control

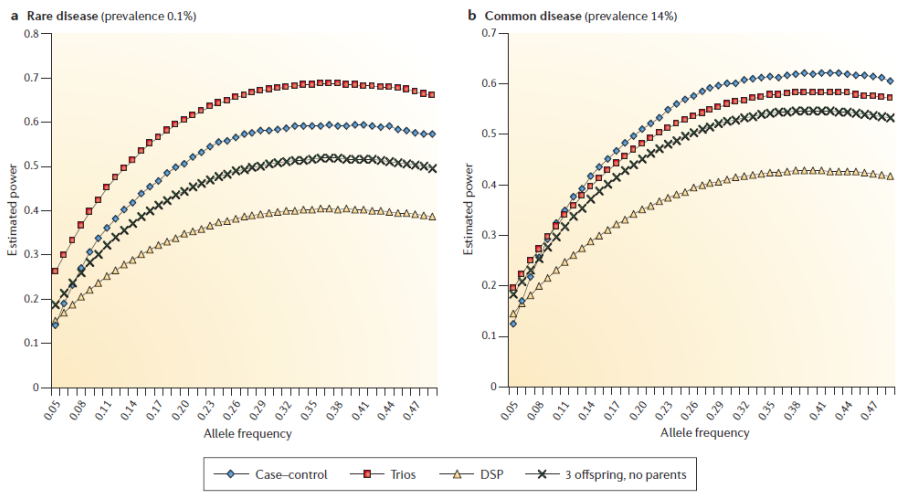


Figure 2-4 Power comparison between case-control studies and family-based designs. The estimated power levels for a case-control study with 200 cases and 200 controls are compared with those for various family-based designs: 200 trios (of an affected offspring plus parents); 200 discordant sibling (sib) pairs (one affected and one unaffected) without parents; 200 '3 discordant offspring (at least 1 affected, at least 1 unaffected) and no parents' (Laird and Lange 2006).

studies (Figure 2-4). The recruitment of probands and their relatives in family-based association studies usually requires more resources in terms of time and money than that of unrelated subjects in population-based studies. Furthermore, more genotyping might be required for family-based studies, and together these factors have increased the popularity of population designs over family-based studies.

However, unlike population-based studies, family-based designs are robust against population substructure, and significant findings always imply both linkage and association (Laird and Lange 2006). Furthermore, studies that use families offer

a solution to the problems of model building and multiple-hypothesis testing, which are important issues in tests of association, and will become more pressing with the advent of genome-wide association studies.

2.1.3 NGS and Family-based studies

The availability of high-throughput genotyping technologies, coupled with the results of major polymorphism characterization efforts such as the International HapMap Initiative, have made it possible to conduct genome wide association (GWA) studies seeking to identify common variations that are statistically linked with particular diseases.

Even though hundreds of GWA studies have been performed and identified statistically compelling associations between particular genetic variations and diseases, many investigators have gone beyond survey sequencing of human genes to catalog rare sequence variations. The purpose of doing this is to actually contrasting and comparing the frequency of rare variations in individuals with and without disease. One important aspect of these studies is that they focused on the identification of rare variations, any one of which might often be possessed by the individuals with the disease phenotype of interest.

This suggests that once one collected the homozygous samples, then no matter how small samples are collected, it is possible to identify all the variations contributing to the variations that were greater in frequency among individuals. A good example consists of homozygous status is the family data. Thus I make use of

benefits of homogeneous in a family, together with sequencing technology, to analyze causal variants with small sample data.

2.2 Overview on Existing Approaches with Small Samples

The most of genomic data analyses have been performed with large amount of samples. Genome-wide association (GWA) studies are one of examples. In GWA studies, more than thousands of case and control samples are in need to obtain meaningful results. Thus it is reasonable for GWA studies be left in the field of common variants analysis measured by microarray platform. NGS, however, suits for small samples with rare variants analysis.

Previous analyses with such data (small samples with rare variants) have been conducted as follows (Ng, Bigham et al. 2010, Ng, Buckingham et al. 2010, Ng, Nickerson et al. 2010):

1. variant calling
2. filtering by quality scores
3. screening by various criteria

In screening step, common variants, synonymous mutation, SIFT and PolyPhen scores (Flanagan, Patch et al. 2010), and etc. were used to reduce the number of candidate causal variants associated to the phenotype of interest. However, a typical human genome differs from the reference genome at over 10,000 potentially functional sites; identifying the disease-causing mutation among this plethora of variants can be a significant challenge. For this reason, exome sequencing is preceded by genetic linkage analysis, which allows variants outside of linkage peaks to be excluded. This combination strategy has been successfully used to identify

variants causing autosomal dominant (Johnson, Mandrioli et al. 2010, Wang, Yang et al. 2010) and recessive diseases (Bilguvar, Ozturk et al. 2010, Bolze, Byun et al. 2010, Otto, Hurd et al. 2010, Sirmaci, Walsh et al. 2010, Walsh, Shahin et al. 2010, Abou Jamra, Philippe et al. 2011, Kalay, Yigit et al. 2011), as well as those affecting quantitative traits (Bowden, An et al. 2010, Musunuru, Pirruccello et al. 2010, Rosenthal, Ronald et al. 2011).

There are two alleles in the two homologous chromosomes at a certain marker. One of the two alleles received from one of the two alleles of father with equal probability, and likewise one of the two allele from one of the two alleles of mother with equal probability. If an individual with genotype A_1A_2 has the same phenotype as the individual with genotype A_1A_1, but has different phenotype from the individual with genotype A_2A_2, then allele A_1 is said to be dominant to A_2, or equivalently, allele A_2 is said to be recessive to allele A_1. In addition, the phenotype associated with A_1 is said to be dominant, and phenotype associated with A_2 is said to be recessive.

If an individual with genotype A_1A_2 has different phenotype with both A_1A_1 and A_2A_2, the allele A_1 and A_2 are said to be codominant. For example, there are two alleles D and d at disease locus. Let D be a disease gene (or allele). If the disease is a dominant disease, an individual with genotype DD or Dd will have the disease. An individual with genotype dd will be normal. If the disease is a recessive disease, an individual with genotype DD will have disease, while an individual with genotype Dd or dd will be normal.

As described in the previous section, one of the best strategies for sequencing data with small samples in homogenous environments. Homogenous environments means data from family or from autosomal dominant Mendelian disease, in which few strong candidates are highly associated with the disease.

2.2.1 Filtering & Screening Analysis

Whole exome sequencing (WES) has become a popular strategy for discovering potential causal variants in individuals with inherited Mendelian disease. Most studies of this strategy adopt a simple analysis, call it a ′filtering & screening analysis′. Let′s take an example of a paper published on nature genetics in 2010 (Ng, Buckingham et al. 2010). They examined four individuals of self-reported European ancestry with Miller syndrome from three unrelated families. In two families, two siblings were affected and in one family a single individual had been diagnosed with Miller syndrome (Figure 2-5).

Genomic DNA was extracted from peripheral blood lymphocytes, and

	Kindred 1-A		Kindred 1-B		Kindred 1 (A+B)		Kindreds 1+2		Kindreds 1+2+3	
Filter	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive
NS/SS/I	4,670	2,863	4,687	2,859	3,940	2,362	3,099	1,810	2,654	1,525
Not in dbSNP129	641	102	647	114	369	53	105	25	63	21
Not in HapMap 8	898	123	923	128	506	46	117	7	38	4
Not in either	456	31	464	33	228	9	26	1*	8	1*
Predicted damaging	204	6	204	12	83	1	5	0	2	0

Each cell indicates the number of genes with nonsynonymous (NS) variants, splice acceptor and donor site mutations (SS) and coding indels (I). Filtering either by requiring the presence of NS/SS/I variants in siblings (kindred 1 (A+B)) or of multiple unrelated individuals (columns) or by excluding annotated variants (rows) identifies 26 and 8 candidate genes under a dominant model and only a single candidate gene, *DHODH*, under a recessive model (light gray cells). Exclusion of mutations predicted to be benign using PolyPhen (row 5) increases sensitivity under a dominant model but excludes *DHODH* under a recessive model because a variant in kindred 1 is predicted to be benign. A single candidate gene is identified in kindred 1 under a recessive model and excluding benign mutations (dark gray cell), but this candidate is excluded in comparisons with unrelated cases of Miller syndrome. Mutations in this candidate, *DNAH5*, were found to cause a primary ciliary dyskinesia in kindred 1. The asterisk indicates that a second gene, *CDC27*, was also identified as a candidate gene, but this is due to the presence of multiple copies of a processed pseudogene that recurrently gave rise to a false positive signal in exome analyses.

Figure 2-5 Direct identification of the gene for a mendelian disorder by exome resequencing (Ng, Buckingham et al. 2010)

performed sequencing. Reads were mapped to the reference human genome (UCSC hg18), initially with efficient large-scale alignment of nucleotide databases (ELAND) software (Illumina) for quality recalibration and then again with Maq. After calling variants from the data, they filtered out the variants by comparing against eight HapMap individuals for whom they had previously reported exome data. They also screened out the variants with PolyPhen score for all nonsynonymous SNPs. Any nonsynonymous variant that was not assigned a ′benign′ PolyPhen prediction was considered to be damaging, as were all splice acceptor and donor site mutations and all coding indels. By this ′filtering and screening′ analysis, they left with a candidate gene, DHODH, which encodes a key enzyme in the pyrimidine de novo biosynthesis pathway. Sanger sequencing confirmed the presence of DHODH mutations in three additional families with Miller syndrome.

2.2.2 Linkage analysis

Linkage analysis provides researchers a powerful method for mapping the location of disease-causing loci by identifying genetic markers that are co-inherited with a phenotype of interest. Usually the LOD score (log of odds), developed by Newton E. Morton, is used for a statistical test for linkage analysis. The LOD score compares the likelihood of obtaining the test data if the two loci are indeed linked, to the likelihood of observing the same data purely by change. Positive LOD scores favor the presence of linkage, whereas negative LOD scores indicate that linkage is less

likely. Briefly, it works as follows:

1. Establish a pedigree
2. Make a number of estimates of recombination frequency
3. Calculate a LOD score for each estimate
4. The estimate with the highest LOD score will be considered the best estimate.

The LOD score is defined as,

$$\text{LOD} = \log_{10} \frac{(1-\theta)^{NR} \times \theta^{R}}{0.5^{(NR+R)}}$$

where, NR and R denote the number of non-recombinant offspring and recombinant offspring, respectively.

2.2.3 Copy number variation (CNV) analysis

Copy number variation (CNV) analysis refers to the process of analyzing data produced by a test for DNA copy number variation in patients′ sample. The analysis detects chromosomal copy number variation that may cause or may increase risks of various diseases. For sequencing data, read depth difference can be understood as the copy number inherited in genomic DNA. Therefore, inferring CNVs by analyzing read depth across the genome and comparing CNVs among the individuals

of interests provides candidate causal variants for the comparison of interest.

2.3 Prioritizing Disease Causing Variants with Deafness Family

In this section, an approach not only for reducing the exome search space but also for making full usage of data for Mendelian diseases using a multiphasic analysis is proposed. A multiphasic analysis is designed of copy number variation (CNV), linkage, and single nucleotide variation (SNV) analysis of whole exome sequencing (WES) data for the efficient discovery of Mendelian disease. Here, the analysis is applied to a nonsyndromic hearing loss (NSHL)-causing mutation data. From whole exome sequencing (WES) data, five distinct CNV loci were identified from a NSHL family WES data, but they were not co-segregated among patients. Linkage analysis based on SNVs selected six candidate loci (LOD >1.5). 15 SNVs that co-segregated with NSHL in this family were selected, which were located in the six linkage candidate loci. Finally, the novel variant p.M305T in ACTG1 (DFNA20/26) was selected as a disease-causing variant. Here a multiphasic approach is introduced to analyze CNV, linkage, and SNV of WES data for the identification of a candidate mutation causing NSHL. This stepwise multiphasic approach enabled us to expedite the discovery of disease-causing variants from a large number of variants in patients.

2.3.1 Background of the study

By virtue of the recent development of massively parallel DNA sequencing technologies, access to genomic composition has become easier than ever. With the

advantage of exome sequencing, many studies have identified causal variants responsible for numerous disorders. Exome sequencing provides a particularly powerful method with which to identify disease-causing single nucleotide variation (SNV) in Mendelian disorders (Musunuru, Pirruccello et al. 2010, Yang, Ahmed et al. 2010, Bamshad, Ng et al. 2011, Min, Kim et al. 2011). Though whole exome sequencing (WES) has successfully resulted in the discovery of many genes that cause Mendelian disorders, the analysis of WES data is still challenging (Bamshad, Ng et al. 2011). Each individual exome has more than 20,000 variants as compared with the reference genome. Even in familial Mendelian disorders, overall success rates in identifying disease-causing genes are around 50% (Gilissen, Hoischen et al. 2012). Many potential reasons are presumed for failure in the WES approach that need to be solved fully for the full promise of WES in routine diagnostics for Mendelian disorders.

Filtering patient data against those from normal populations and inferring identity-by-descent (IBD) in family studies could enrich the candidates (Krawitz, Schweiger et al. 2010, Musunuru, Pirruccello et al. 2010). Genetic linkage analysis has also been a powerful tool to isolate potential causal candidate variants. A two-step approach of linkage analysis using a single nucleotide polymorphism (SNP) microarray to detect high logarithm of odds (LOD) score regions and subsequent targeted re-sequencing of the regions has been utilized in many genomic studies to intensify the power of detection (Cooper and Shendure 2011). Classically, microsatellite markers have been used for linkage analysis, and now millions of

dimorphic SNP markers could provide higher resolution to pinpoint candidate loci (Nielsen, Paul et al. 2011). Currently, there are many efforts to use coding SNP information from WES data to facilitate genetic linkage mapping. Using coding SNP data from WES, I can establish a multiphasic exome analysis based on linkage and SNVs (Musunuru, Pirruccello et al. 2010, Smith, Bromhead et al. 2011).

Copy number variation (CNV) has been implicated in both Mendelian diseases (Cho, Kim et al. 2010) and common traits, such as obesity (Sha, Yang et al. 2009) and schizophrenia (International Schizophrenia 2008). The presence of large insertions or deletions in patients is typically investigated prior to SNV analysis by karyotyping, fluorescence in situ hybridization (FISH), or array comparative genome hybridization (aCGH). Estimation of CNV is a challenging field of WES analysis, in which local depths of coverage must be mapped to copy numbers. aCGH has limitations in detecting high CNV regions. Conversely, CNV data based on WES provide more accurate copy numbers, because depths of exon coverage from WES data vary linearly with real copy numbers (Kang, Gan et al. 2010). Bioinformatics tools to analyze copy numbers from WES data are now publicly available (Li, Lupat et al. 2012).

2.3.2 Background of the nonsyndromic hearing loss (NSHL)

Nonsyndromic hearing loss (NSHL) contributes to more than 70% of inherited hearing loss. Until now approximately 50 genes have been known to relate causally to NSHL. Many studies identified more than 129 loci responsible for NSHL.

However, 47 loci have not yet been mapped to proper genes (Hilgert, Smith et al. 2009, Dror and Avraham 2010). The complexity of the auditory system may be why so many genes and loci are linked to hearing loss. Genetic causes of hearing loss can be detected by sequence analysis, which helps clinicians and patients to delineate the basis of disease. Given that hearing loss in early childhood can affect linguistic development (Dror and Avraham 2010), it is important to improve current techniques in identifying genetic alterations that cause this disease. Earlier identification of such alterations in patients and families may allow for better clinical management of NSHL.

2.4 Materials and Methods

This study was co-worked with many others. I emphasize on experimental works and clinical information were obtained from collaborators.

2.4.1 Subjects

This study was approved by Institutional Review Boards (IRBs) at the Seoul National University Hospital (SNUH) and the Seoul National University Bundang Hospital (SNUBH). The written informed consent for participation in the study was obtained from participants or, where participants are children, a parent or guardian. A three generation pedigree was established for the family (SNUH3) (Figure 2-6). Among 15 subjects in the SNUH3 family, 13 subjects were willing to participate in

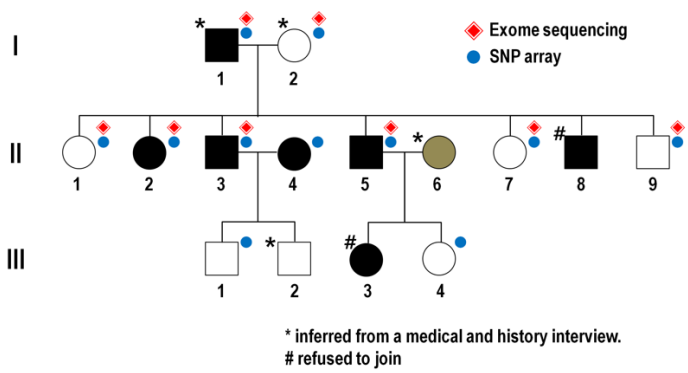


Figure 2-6 Pedigree with information of phenotype and performed experiments

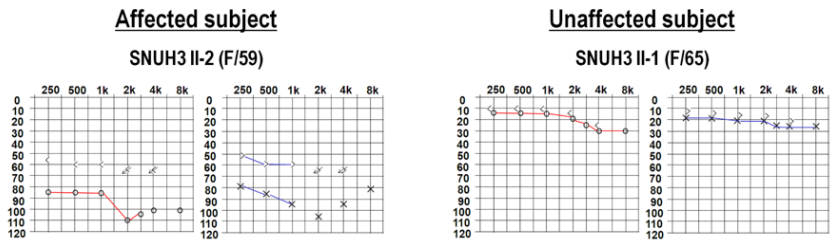


Figure 2-7 Typical audiograms of affected and unaffected subjects

this genetic study, but two reportedly deaf subjects (II-8 and III-3) refused participatioin. DNA from blood lymphocytes was isolated from 12 subjects, and DNA from III-2 was obtained through a buccal swab.

2.4.2 Audiometric analysis

Pure tone and speech audiometry and physical examinations were performed in nine members of the cohort (Figure 2-7). Pure tone audiometry (PTA) with air and bone conduction at frequencies ranging from 250 to 8,000 Hz was obtained from the recruited subjects according to standard protocols. The auditory phenotype was inferred from the thorough medical and developmental history interviews from one deaf subject (I-1), two likely unaffected subjects (I-2 and III-2), and one subject (II-6) with an equivocal hearing status.

2.4.3 Whole exome sequencing (WES)

Eight of 13 recruited subjects (four affected and four unaffected) were chosen for commercial WES (Otogenetics, Norcross, GA) and analyzed as previously reported (Min, Kim et al. 2011). Briefly, paired-end reads of 100 bp from the eight subjects were aligned by bwa-0.6.1 with default settings on UCSC hg19 reference genome. SAMtools (Li, Handsaker et al. 2009) and Picards were used to process SAM/BAM files and mark duplicates. Local realignment around indels and base quality score recalibration were done for each sample, and variants were called by a unified genotyper in GATK-1.3 (McKenna, Hanna et al. 2010). Perl script and ANNOVAR (Wang, Li et al. 2010)were used to annotate variants and search the relevant known SNPs and indels from dbSNP135 and the 1000 Genome database. Variants with a read depth greater than 10 and genotype quality score greater than 30 were filtered in to analyze in the next steps.

2.4.4 CNV analysis using WES data

CNVs were detected by CONTRA software (Li, Lupat et al. 2012) that uses BEDTools to calculate coverage per exon and applies statistics to normalize coverage data and test the fold changes. A new baseline file was produced using our data, but distinct deletions or amplifications were expected to be detected. Polymerase chain reaction (PCR) duplicates were removed by Picards before running CONTRA.

We tabulated a 3×2 exon copy variation contingency table based on the whole per-exon CNV status of eight subjects (Table 2-1). We conducted a Fisher exact test

to assess the significance of the differences between proportions of abnormal copy number events occurring in affected and in unaffected family members. We assumed all the subjects were independent to conduct an alternative practical method to find loci that segregated with disease.

Table 2-1 Copy number variation contingency table

Copy number	**Patient**	**Normal**	**Total**
gain	n_{21}	n_{20}	n_{2+}
normal	n_{11}	n_{10}	n_{1+}
loss	n_{01}	n_{00}	n_{0+}
	n_{+1}	n_{+0}	8

2.4.5 Linkage analysis using WES and SNP microarray

From WES data, the following variants were filtered out; those located on sex chromosomes, those with low coverage (<10X), and those with a low genotype quality score (<30) in any of the eight subjects with 17,498 SNVs. Genome-Wide Human SNP Array 6.0 (Affymetrix, Santa Clara, CA), which contains 328,125 SNP markers in autosomal chromosomes was used. I performed parametric linkage analysis with R package paramlink. The pedigree suggests an autosomal dominant mode of inheritance, so I assumed an autosomal dominant model with default values of full penetrance (f0, f1, f2) = (0, 1, 1) and disease allele frequency = 1e-05. Here,

the penetrance parameters f0, f1, and f2 are also defined using conventional notation as below.

$$f_i = P(\text{affected}|i \text{ copies of the disease allele})$$

The recombination fraction between the disease locus and the marker was set to θ=0 by default. Single-point LOD scores for all the markers were computed and LOD scores from SNP microarray and WES were compared. The subjects and the markers that appeared in both platforms in common were matched using manual python and R scripts. Finally, single-point analyses were performed with all data.

2.5 Results

2.5.1 Clinical features of a NSHL family

We identified a Korean family with six members affected with NSHL and seven unaffected members (Figure 2-7). Pure tone audiometry (PTA) was performed on nine family members, and three members (II-2, II-3, and II-5) exhibited profound post-lingual hearing loss. The three members had normal cognitive function and no anomalous-looking features. They went through a battery of clinical tests from general physical examinations, chest x-ray and simple blood tests to detailed imaging studies such as brain MRI and temporal bone CT. No abnormality was detected in the tests, excluding a possibility that hearing loss in these patients is syndromic. The other siblings (II-1, II-7, and II-9) demonstrated normal hearing (Figure 2-8). Patients II-2, II-3, and II-5 estimated that their hearing loss became severe in their 30's and started to wear hearing aids. Their hearing loss further aggravated and became profound in their late forties. Finally, patients II-3 and II-5 eventually did not benefit from the hearing aid any longer and underwent cochlear implantation. They achieved recognition of common sentences without lip reading one year after implantation. We investigated the sequence of the GJB2 in the patients, which is one of the most frequently detected genes in individuals with NSHL. Because there was no mutation in GJB2, we performed WES on several members of this family to identify a disease-causing mutation.

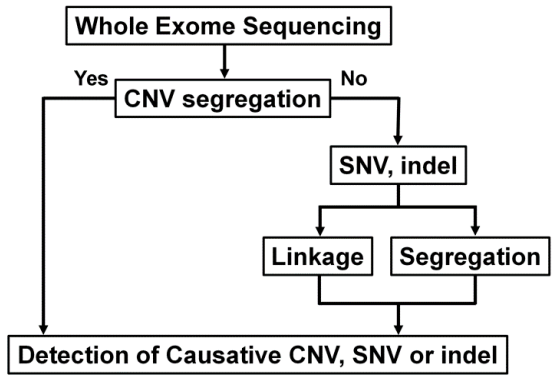


Figure 2-8 A scheme of multiple parallel analysis of WES data

2.5.2 Copy number analysis using WES data

WES data were analyzed in parents and six siblings (four affected and four unaffected members, Figure 2-7). Mean coverage of each sample ranged from 40.3X to 51.3X, and 87.0% to 90.5% of the targeted exome had at least 10 reads. A multiphasic WES analysis was designed to find a causative NSHL mutation (Figure 2-8). First, we investigated co-segregation of copy number duplication or deletion in exomes of patients using CONTRA software. We detected five loci with CNVs with distinct features in the plots (Figure 2-9A). None of these CNVs co-segregated with affected or unaffected family members. One locus showing the CNVs from three members (high copy number exons in II-9, and low copy number exons in II-3 and II-7) was located in 8p23.1, a region that contains the beta-defensin genes and SPAG11 (Figure 2-9B). The following genes were identified to be located at the

regions of distinct CNVs in the indicated family members; GSTM1 in 1p13.3 (I-1, II-3, II-7, and II-9) (Figure 2-9C), UGT2B17 in 4q13.2 (I-2 and II-7), BNTL3 in 5q35.3 (II-1), and LILRB2 in 19q13.4 (I-2) (data not shown). We also applied a Fisher exact test for the LOD score per exon to detect co-segregated regions of CNVs, but there was no peak with a value reaching significance. We composed two groups defined by the pattern of segregation of SPAG11, GSTM1, and the beta-defensin genes to validate the relevance of this method (Figure 2-9D).

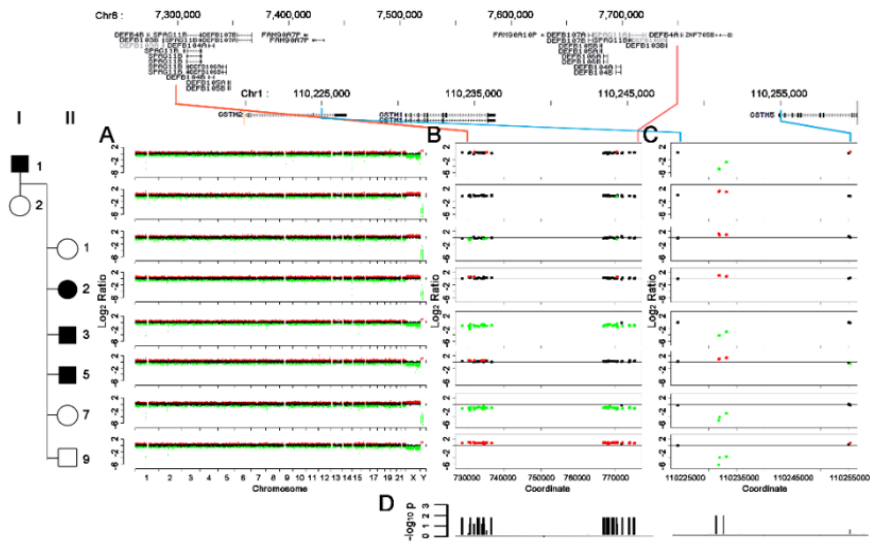


Figure 2-9 CNV detected by WES. CNV throughout the chromosomes – 1p13.3, 5q35.3, 8p23.1, and 19q13.4 have distinct CNVs (14q32.3 is distinct, but contains variable regions associated with antibody production) (A), 8p23.1 containing beta-defensin clusters (B), and 1p13.1 containing *GSTM1* (C) of eight subjects. Red and green dots are exons with $p<0.05$.

2.5.3 Exome linkage analysis

We used 17,498 coding autosomal SNVs from WES data and performed single-point linkage analysis, because the pedigree strongly suggests an autosomal dominant mode of inheritance. We identified six "hot spots" where a number of peaks were closely clustered (Figure 2-10). Each peak at chromosome 3, 11, 13, 14, 16, and 17 consists of 11, 67, 2, 13, 17, and 13 exons, respectively.

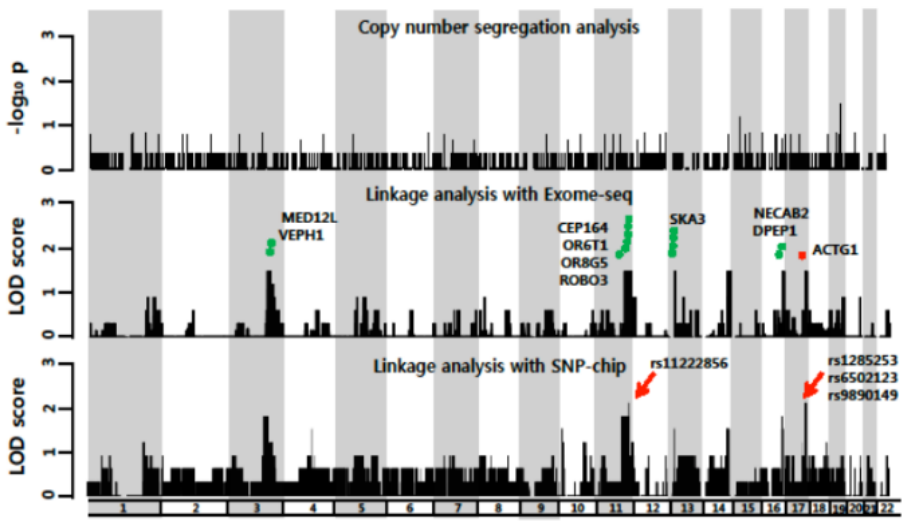


Figure 2-10 A multiphasic analysis of WES data. WES data were analyzed for exon CNVs and SNVs. Fisher exact test detected one exon segregating with NSHL on chr19 (A). Linkage analysis with SNVs called by Exome-seq identified six disease-linked "hot spots" on chr3, chr11, chr13, chr14, chr16, and chr17 (B). Segregation analysis independently identified 15 SNVs co-segregating with NSHL (green dots). Among them, a novel variant resulting in p.M305T, in ACTG1 on chr17 was validated with Sanger sequencing (red dots). Linkage analysis was also performed with SNP microarray by adding three more subjects in the family. Not only were similar "hot spots" detected, adding more subjects in the analysis enhanced the true peak (red arrow) (C).

We validated single-point linkage using a SNP microarray containing 328,125 SNPs. Along with the initial eight family members recruited for WES analysis, we included three additional subjects (II-4, III-1, and III-4) for validating the significance of peaks obtained from exome linkage analysis. The six "hot spots" detected with sequencing data were also detected in microarray analysis with relatively high LOD score (Figure 2-10). Adding three more subjects into the linkage analysis enhanced the peaks at chromosome 11 and 17. These two points consist of one and three peaks (LOD score >2) at chromosome 11 and chromosome 17, respectively. Genotype patterns of these four peaks were perfectly matched with an autosomal dominant mode of inheritance.

2.5.4 SNV analysis

From the WES analysis of four affected and four unaffected family members,

Table 2-2 Nonsynonymous SNVs and indels identified in patients but not in non-symptomatic family members

Gene	Chr	Nucleotide variation	Amino acid variation	Frequency in 1,000 genome	dbSNP135
MED12L	chr3	c.G3629A	p.R1210Q	0.23	rs3732765
VEPH1	chr3	c.T1564C	p.S522P	0.28	rs11918974
CWF19L2	chr11	c.A2681G	p.Y894C	0.27	rs3758911
CEP164	chr11	c.G281A	p.S94N	0.19	rs490262
OR6T1	chr11	c.G465C	p.W155C	0.0046	rs150534954
OR8G5	chr11	c.G287A	p.C96Y	0.45	rs2512168
OR8G5	chr11	c.G716A	p.G239E	0.50	rs2512167
ROBO3	chr11	c.G1247A	p.R416H	0.14	rs3862618
SKA3	chr13	c.A1157G	p.K386R	0.13	rs11147976
SKA3	chr13	c.C1142T	p.T381I	0.11	rs11147977
SKA3	chr13	c.G559A	p.V187I	0.14	rs61950353
SKA3	chr13	c.208delC	p.Q70fs	-	rs151272242
NECAB2	chr16	c.C704G	p.T235S	0.20	rs2292324
DPEP1	chr16	c.G1051C	p.E351Q	0.24	rs1126464
ACTG1	chr17	c.T914C	p.M305T	-	-

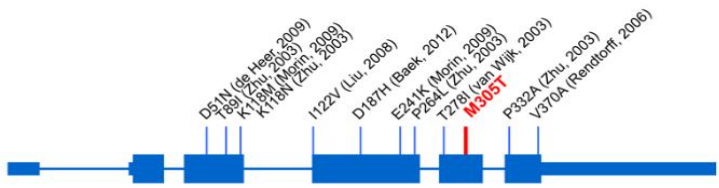


Figure 2-11 The p.M305T (black text) mutation reported in this text and several other previously reported mutations (red(Zhu, Yang et al. 2003), orange(van Wijk, Krieger et al. 2003), green(Liu, Li et al. 2008), purple(de Heer, Huygen et al. 2009), and cyan(Morin, Bryan et al. 2009)) in ACTG1 protein cause hearing loss

we identified 18,748~20,025 SNVs and 413~457 indels. These were reduced to 962~1,123 SNVs and 140~153 indels after filtering through dbSNP135 and 1000 Genome databases. Fifteen variants causing amino acid changes were selected based on their co-segregation pattern within the family (Table 2). All of the 15 variants in chromosomes 3, 11, 13, 16, and 17 corresponded to regions of high LOD scores

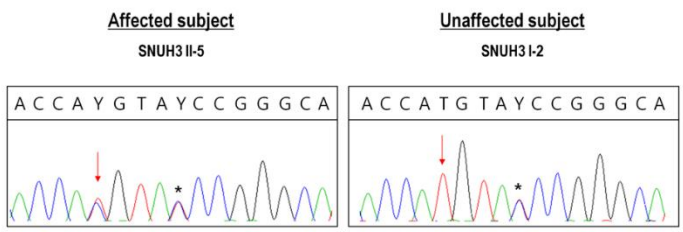


Figure 2-12 p.M305T, confirmed by Sanger sequencing, perfectly co-segregates with hearing loss

(Figure 2-10). One novel mutation was a methionine to threonine change at amino acid 305 (p.M305T) in actin, gamma 1 (ACTG1). This candidate variant was validated by Sanger sequencing, and co-segregated with hearing loss in all family members (Figure 2-11 and Figure 2-12).

ACTG1 (DFNA20/26; MIM: 604717) is a strictly conserved gene among nineteen out of twenty eukaryotes (HomoloGene:74402) and the M305 codon is conserved in the nineteen species. Protein damage prediction analysis termed this mutation as "possibly damaging" by HumDiv, "probably damaging" by HumVar in Polyphen2 (Adzhubei, Schmidt et al. 2010), and "disease causing" by MutationTaster (Schwarz, Rodelsperger et al. 2010). The mutation site, Met305, was visualized in the 3D structure of bovine beta-actin bound by ATP with profilin

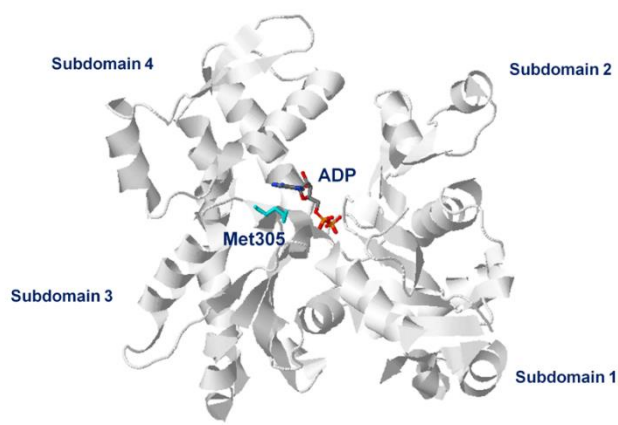


Figure 2-13 Methionine located in the vicinity of bound ADP (yellow: Met305, molecule: ADP)

(Figure 2-13). The methionine was located beside the ATP. Additionally, Met305 was listed as one of the predicted residues in binding sites for ATP by Protein Data Bank (PDB).

2.5 Conclusion & Discusion

WES is a powerful technique to discover causative genes in human diseases. Although it has been integral in identifying more than 1,000 novel genes in Mendelian disorders (Bamshad, Ng et al. 2011), there is still a need for an increase in the efficiency of gene discovery using WES data. In this regard, we analyzed WES data from a family with a history of NSHL focusing on three categories of genetic information: CNV, linkage, and SNVs. Utilizing these data, we undertook a stepwise multiphasic approach to identify the disease-causing variations in this family.

8p23.1, containing a beta-defensin cluster, was detected as a region with high copy number (II-9) and low copy number (II-3 and II-7) (Figure 2-9). The defensin cluster, containing both alpha- and beta-defensins, was previously studied as a dynamic genomic region with varying copy numbers ranging from one to twelve (Hollox, Barber et al. 2008). The parents had normal copy numbers, which were in contrast to low copy numbers seen in two children, and a high copy number observed in one child. Four total haplotypes of 8p23.1 may be inherited in this family, and each parent may have both under- and over-amplified alleles of 8p23.1. The overall copy number of each parent can appear to be normal due to compensation of copy number from over- to under-amplified alleles (Carelle-Calmels, Saugier-Veber et al. 2009). GSTM1, UGT2B17, genes with frequent reported deletions (Huang, Chen et al. 2009), BNTL3 and LILRB2 also had CNVs in this family. We performed a Fisher exact test on the affected and unaffected family members after validating this method

on the 8p23.1 and GSTM1 groups to determine the amplification or deletion of multiple exons that matched the co-segregation pattern of disease. Multiple statistically significant peaks at 8p23.1 and GSTM1 were identical to plots from the first approach. However, there was only one statistically significant peak identified by testing the two groups that segregated with disease, and this peak did not match the disease status. In this sense, WES may provide a method to identify CNV regions with highly similar sequences, although determining accurate copy numbers may prove difficult.

Linkage analysis was performed to narrow down the number of candidates based on the WES data. Linkage analysis with a relatively small number of markers still provides useful information. Fewer markers from WES data are available than can be obtained from a SNP microarray. Those markers that are identified are not evenly distributed. Given these limitations, it is necessary to consider the potential disadvantages of this approach. Because we analyzed only exonic SNPs (~1% of genome-wide SNPs) we might have lost critical information from outside of exons. Additionally, potential genotyping errors in linkage analyses may reduce power to detect linkage peaks or result in false positive linkage peaks (Cherny, Abecasis et al. 2001). Even so, the results obtained from different data sets in this study confirmed the validity of our approach. Note also linkage analysis needs a large number of subjects in helping to identify putative loci. Unless a proper number of subjects are available, no informative result will be easily obtained.

Co-segregated variants were all located in the loci of high LOD scores after

applying linkage analysis results. However, linkage analysis can decrease the number of candidate variants, particularly in instances in which candidate variants are widely distributed. Additional linkage analysis of WES data demonstrated a similar performance to that of SNP microarray data and generated the result simultaneously with variant calling. Considering that CNVs could be also detected, this multiphasic analysis of WES results efficiently narrowed and identified candidate variants and was advantageous to established methods such as initial aCGH, variant calling only by WES, or prior linkage analysis by SNP microarray.

Actin is a highly conserved cytoskeletal protein that plays important roles in eukaryotic cell processes such as cell division, migration, endocytosis, and contractility. Actin isoforms are classified into two groups based on expression patterns. ACTA1, ACTA2, ACTC, and ACTG2 are “muscle” actins, the predominant isoforms in striated orand smoothcardiac and skeletal muscle, whereas ACTB and ACTG1 are cytoplasmic “non-muscle” actinsthe predominant isoforms found in allnon-muscle cells (Sonnemann, Fitzsimons et al. 2006). Autosomal dominant progressive sensorineural hearing loss, DFNA20/26 (MIM: 604717), is caused by a mutation in the gamma-actin gene on chromosome 17 at q25.3. Some ACTG1 mutations are associated with Baraitser-Winter syndrome which is characterized by developmental delay, facial dysmorphologies, brain malformations, colobomas and variable hearing loss. The constellation of these abnormalities is suggested as the most severe phenotype of ACTG1 mutations (Drummond, Belyantseva et al. 2012, Riviere, van Bon et al. 2012). A genome-wide screen of DFNA20 localized

candidates to 17q25.3 (Morell, Friderici et al. 2000) and mapped the causative missense mutations to highly conserved actin domains in the gamma-actin gene (ACTG1) (van Wijk, Krieger et al. 2003, Zhu, Yang et al. 2003). In vivo and in vitro studies of ACTG1 revealed that it is required for reinforcement and long-term stability of actin filamentous structures of stereocilia but not for auditory hair cell development, which is in line with the progressive nature of the hearing loss related to ACTG1 mutations in humans (Belyantseva, Perrin et al. 2009, Perrin, Sonnemann et al. 2010). Interestingly, 11 mutations causing DFNA20 (van Wijk, Krieger et al. 2003, Zhu, Yang et al. 2003, Rendtorff, Zhu et al. 2006, Liu, Li et al. 2008, de Heer, Huygen et al. 2009, Baek, Oh et al. 2012) and 6 mutations causing Baraitser-Winter syndrome that have been reported in this gene are all missense mutations. The predicted interaction between Met305 and ATP in bovine beta-actin, a protein with 99% identity of ACTG1, implies that the mutation of Met305 may influence adenosine triphosphate (ATP) binding of ACTG1, which is essential for polymerization from G-actin to F-actin.

ACTG1 is predominantly expressed in intestinal epithelial and auditory hair cells (Khaitlina 2001). Detection of exclusively missense mutations in this gene may imply that truncating mutations cause more severe effects that might cause embryonic lethality. Our hearing impaired subjects (II-2, II-3, and II-5) denied any gastrointestinal complaint. The subjects in this study required cochlear implants, recapitulating what has previously been reported regarding the management of patients with mutations in ACTG1 and resultant NSHL (Rendtorff, Zhu et al. 2006).

A severe phenotype and rapid progression of hearing loss to a profound level within one or two decades associated with mutations in this gene necessitates an early molecular genetic diagnosis and timely auditory rehabilitation.

Two or more platforms (aCGH, SNP array, and WES) have previously been needed to generate complex genetic information such as CNV, linkage, SNV and indels. WES has primarily been utilized to obtain only SNVs and indels in general studies of Mendelian disorders. Our study agrees well with other work in demonstrating that WES data analysis allows for CNV and linkage determination by means of its quantitative traits. Given the robust nature of WES data, it is clear that the full capabilities of this relatively new technology have not yet been realized. Our multiphasic WES analysis proved very powerful in the interpretation and narrowing of WES results, in particular when a large number of family data is available.

CHAPTER 3

TRNASCRIPTOMIC DATA ANALYSIS

This chapter focuses on differential expression (DE) analysis of RNA-Seq experiments with a small number of replicates. To do that, overview of the RNA-Seq analysis will be summarized after providing some characteristics of RNA-Seq. The existing methods for DE analysis are reviewed and the proposed method is introduced. Both real and simulated data analyses will follows.

3.1 Introduction

RNA-Sequencing (RNA-Seq) technique provides a valuable information of characterizing the molecular nature of the cells. Unfortunately, expense and limited specimen availability often lead to studies with small sample sizes and hypothesis testing on differential expression between classes is generally difficult. The problem is especially challenging with non-replicated data because estimation of variability is not feasible. Thus most of existing methods for this problem are based on a need for replicates. In this thesis, I introduce a simple but robust method with local pooling and outlier removing to account for non-replicated RNA-Seq data. For the discovery of differential expression of two classes, the new method allows examination of non-replicated RNA-Seq experiments. The validity of the method is demonstrated using both real and simulated datasets. By comparing the results made by the proposed method to those from others, I found that the proposed method, in general, performs better than other competing methods with a small number of replicates in various simulation settings; it shows consistently high true discovery rate, while not increasing the rate of false positives.

3.1.1 From microarray to RNA-Seq

Several comparisons of RNA-Seq and microarray data have been made. These include proof-of-principle demonstrations of the sequencing platform, dedicated comparisons studies and analysis methodology development. The results are

unanimous: sequence has higher sensitivity and dynamics range, coupled with lower technical variation. Furthermore, comparisons have highlighted strong concordance between microarrays and sequencing in measures of both absolute and differential expression. Nevertheless, microarrays have been, and continue to be, highly successful in interrogating the transcriptome in many biological settings (Tempfer, Riener et al. 2004, Modlich, Prisack et al. 2005, Ruano, Mollejo et al. 2010, Gonzalez-Navarro and Belanche-Munoz 2011).

Microarrays and sequencing each have their own specific biases that can affect the ability of a platform to measure DE. It is well known that cross-hybridization of microarray probes affects expression measures in a non-uniform way and sequence content influences measured probe intensities. Meanwhile, several studies have observed a GC bias in RNA-Seq data and RNA-Seq can suffer from mapping ambiguity for paralogous sequences (Risso, Schwartz et al. 2011).

However, several additional factors (including larger dynamic range and sensitivity) have contributed to the rapid uptake of sequencing for DE analysis. First, microarrays are simply not available for many non-model organisms. By contrast, genomes and sequence information are readily available for thousands of species. Moreover, even when genomes are not available, RNA-Seq can still be performed and the transcriptome can still be interrogated. Second, sequencing gives unprecedented detail about transcriptional features that arrays cannot, such as novel transcribed regions, allele-specific expression, RNA editing and a comprehensive capability to capture alternative splicing.

Sequencing is not without its challenges, of course. The cost of the platform may be limiting for some studies. However, with the expansion in total sequencing capacity and the ability to multiplex, the cost per sample to generate sufficient sequence depth will soon be comparable to that of microarrays.

3.1.2 Overview on RNA-Seq analysis

An overview of the typical RNA-Seq pipeline for DE analysis is outlined in Figure 3-1. First, read are mapped to the genome or transcriptome. Second, mapped read for each sample are assembled into gene-level, exon-level or transcript-level expression nummaries, depending on the aims of the experiment. Next, the summarized data are normalized in concert with the statistical testing of DE, leading to a ranked list of genes with associated P-values and fold changes. Finally, biological insight from these lists can be gained by performing systems biology approaches, similar to those performed on microarray experiments.

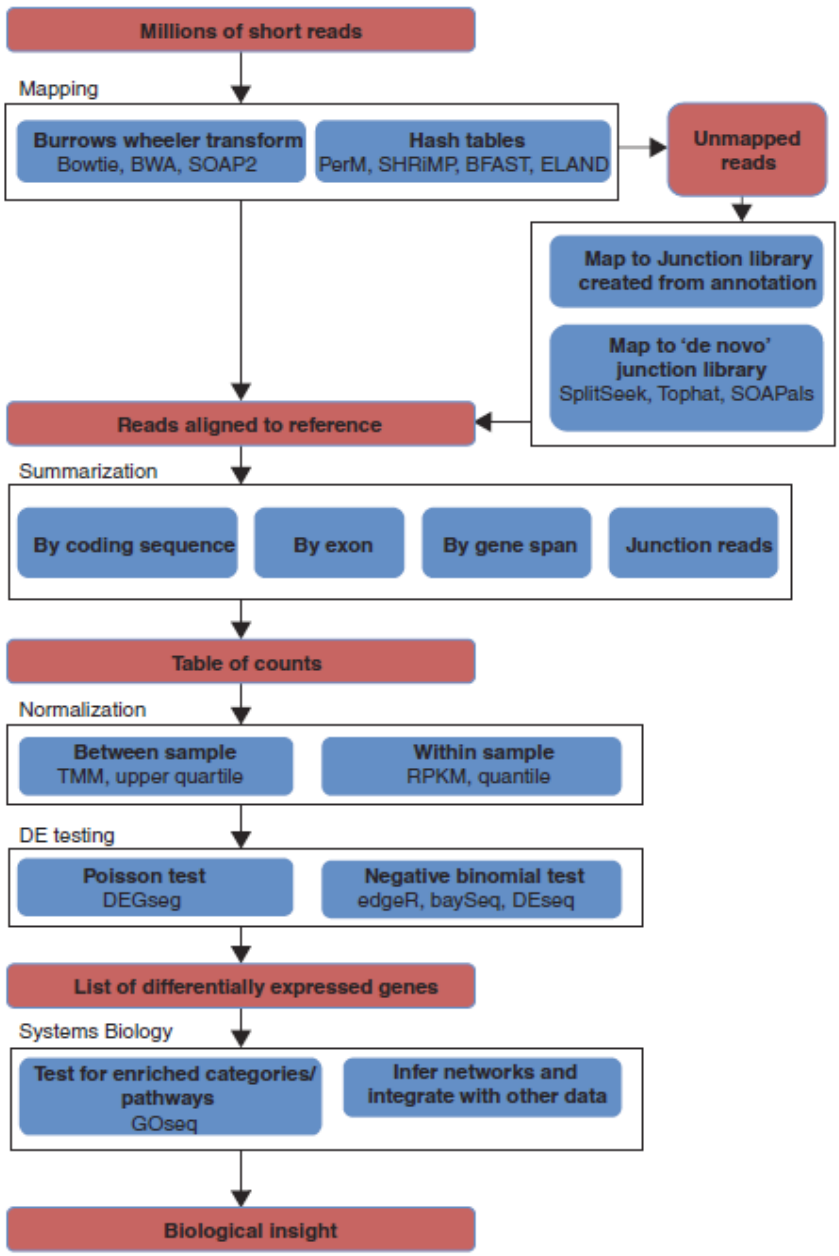


Figure 3-1 Overview of the RNA-seq analysis pipeline for detecting differential expression. From (Oshlack, Robinson et al. 2010)

3.1.3 Differential expression (DE) analysis with small samples

High-throughput sequencing of cDNA that has been derived from an RNA sample, known as RNA-Seq, has recently been developed and applied to various studies depending on the scientific interests: detecting fusion genes, transcribed single nucleotide polymorphisms (SNPs), DE, and so on (Maher, Kumar-Sinha et al. 2009, Gregg, Zhang et al. 2010). In particular, profiling gene expression and testing DE between classes have been the primary process of most biological studies.

The main purpose of a DE analysis is to identify transcripts that have changed significantly in abundance across experimental conditions. This goal has been achieved by many different statistical methods for data in array-based technology. Compared to microarray, however, RNA-Seq has different characteristics, such as a more dynamic range and a lower background expression level measured in count. In order to account for those properties, several methods have been proposed and they modelled RNA-Seq count data using Poisson and negative binomial distribution in the presence of replicated samples (Robinson and Smyth 2007, Marioni, Mason et al. 2008, Anders and Huber 2010, Robinson, McCarthy et al. 2010, Robinson and Oshlack 2010, Wang, Feng et al. 2010, Di, Schafer et al. 2011, Tarazona, Garcia-Alcalde et al. 2011).

A recent review presented a comparative analysis on eleven methods for DE analysis of RNA-Seq data (Soneson and Delorenzi 2013). The review reported that all methods perform well with large sample sizes, while many of them imposed

problems with small sample sizes, such as exceeding false discovery rates. Despite decreasing sequencing costs, RNA-Seq experiments still remain expensive. Moreover, limited specimen availability often lead to studies with a small number of (sometimes none of) replicates. Therefore, the problem of detecting DE with small samples still attracts research interests, especially when no replicates are available. The difficulties in this problem lies in a reliable hypothesis testing based on accurate estimating transcript-specific variance.

3.2 A Review of Existing Methods

Among the methods, edgeR (Robinson, McCarthy et al. 2010, Robinson and Oshlack 2010), DESeq (Anders and Huber 2010), and NBPSeq (Di, Schafer et al. 2011) were reported to have the best results for the small sample size. edgeR assumed that mean and variance are related with a single proportionality constant that is the same throughout the experiments (Robinson and Oshlack 2010). Hence only one parameter needs to be estimated for each transcript, allowing application to experiments with a small number of replicates. DESeq decomposed variance into two terms, a shot noise term and a raw variance term (Anders and Huber 2010). By assuming that the per-transcript raw variance is a function of the expectation value of gene concentration and condition, they extended the model proposed by EdgeR. NBPSeq uses an over-parameterized version of the negative binomial called the NBP distribution, which incorporates a non-constant dispersion parameter directly within a parametric family (Di, Schafer et al. 2011). These methods are based on modelling dispersion of negative binomial distribution and need replicates to estimate variance in order to test for DE, except DESeq.

Two methods are applicable to RNA-Seq data with a non-replicated set: DESeq and NOISeq. NOISeq introduced two differential expression statistics; the log ratio (fold change) and the absolute value of difference (Tarazona, Garcia-Alcalde et al. 2011). Instead of estimating variance, they generated noise (probability) distribution in an empirical way from these two values. A test for DE is then performed using the

odds of the probabilities between two conditions with a specific threshold. However, even though these two methods can be applicable to non-replicated experiments, which these methods are not primarily designated for, more caution should be taken to the assumptions they have made. DESeq assumed that a minority of DE transcripts will not have a severe impact on the gamma-family GLM fit and a valid mean-variance relationship can be estimated from treating the two samples as if they were replicates. However, this assumption is not the most likely especially in the presence of a large number of outliers. NOISeq generates simulated read counts which follow a multinomial distribution and performs a DE test based on these simulated replicates, which might ignore transcript-specific variability.

3.2.1 edgeR

edgeR assume the data can be summarized into a table of counts, with rows corresponding to genes (or tags or exons or transcripts) and columns to samples. For RNA-Seq experiments, these may be counts at the exon, transcript or gene-level. edgeR model the data as negative binomial (NB) distributed,

$$Y_{ij} \sim NB\left(M_k p_{ij}, \phi_i\right),$$

for gene i and sample j. Here, M_k is the library size (total number of reads), ϕ_i is the dispersion and p_{ij} is the relative abundance of gene ij in experimental group k to which sample j belongs.By using the NB parameterization where the mean is $\mu_{ij} = M_k p_{ij}$ and variance is $\mu_{ik}(1 + \mu_{ik}\phi_i)$. For DE analysis, the parameters of interest are p_{ij}.

edgeR estimates the gene-wise dispersions by conditional maximum likelihood, conditioning on the total count for that gene. An empirical Bayes procedure is used to shrink the dispersions towards a consensus value, effectively borrowing information between genes. Finally, differential expression is assessed for each gene using an exact test analogous to Fisher′s exact test, but adapted for over-dispersed data.

3.2.2 DESeq

DESeq is a kind of extension of edgeR. In case of edgeR, mean and variance are related by $\sigma_{gi}^2 = \mu_{gi}(1 + \mu_{gi}\phi_g)$, with a single proportionality constant ϕ_g. However, DESeq are based on three assumptions.

First, the mean parameter μ_{ij}, that is, the expectation value of the observed counts for gene i, in sample j, is the product of a condition-dependent per-gene value $q_{i,\rho(j)}$ and a size factor s_j,

$$\mu_{ij} = q_{i,\rho(j)}s_j.$$

Second, the variance σ_{ij}^2 is decomposed into two terms: shot noise and raw variance term,

$$\sigma_{ij}^2 = \mu_{ij} + s_j^2 v_{i,\rho(j)}.$$

Third, they assume that the per-gene raw variance parameter $v_{i,\rho}$ is a smooth function of q_i, ρ,

$$v_{i,\rho(j)} = v_\rho\left(q_{i,\rho(j)}\right).$$

They also assume that the number of reads in sample i that are assigned to gene g can be modeled by a negative binomial (NB) distribution,

$$K_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2).$$

To estimate the size factors, they rake the median of the ratios of observed counts. Generalizing the procedure just outline to the case of more than two samples.

$$\widehat{s_j} = median_i \frac{k_{ij}}{(\Pi_{v=1}^{m} k_{iv})^{\frac{1}{m}}}.$$

To estimate $q_{i\rho}$, they use the average of the counts from the samples j corresponding to condition ρ, transformed to the common scale:

$$\widehat{q_{i\rho}} = \frac{1}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{k_{ij}}{\widehat{s_j}},$$

Where m_ρ is the number of replicates of condition ρ and the sum runs over these replciates.

First calculate sample variances on the common scale

$$w_{i\rho} = \frac{1}{m_\rho - 1} \sum_{j:\rho(j)=\rho} \left(\frac{k_{ij}}{\widehat{s_j}} - \widehat{q_{i\rho}} \right)^2$$

and define

$$z_{i\rho} = \frac{\widehat{q_{i\rho}}}{m\_\rho} \sum_{j:\rho(j)=\rho} \frac{1}{\widehat{s_j}}.$$

It can be shown that $w_{i\rho} - z_{i\rho}$ is an unbiased estimator for the raw variance parameter by computing the expectation value of $\widehat{w}_{i\rho}$. To simplify notation, the indices i and ρ are dropped in the following. Furthermore, differences between

the true library size s_j and their estimates $\hat{s}_j$ are ignored. Then,

$$\hat{q} = \frac{1}{m}\sum_{j=1}^{m}\frac{K_j}{s_j}$$

is an unbiased estimator of q, because, due to $\mathbb{E}K_j = s_j q_0$. Next,

$$(m-1)\widehat{w} = \sum_{j:\rho_j=\rho}\left(\frac{k_j}{s_j} - \hat{q}\right)^2.$$

For DE testing, i.e., test the null hypothesis $q_{iA} = q_{iB}$, where q_{iA} and q_{iB} are the expression strength parameter for the samples of condition A and B, respectively. To this end, they define, as test statistic, the total counts in each condition,

$$K_{iA} = \sum_{j:\rho(j)=A} K_{ij}, \quad K_{iB} = \sum_{j:\rho(j)=B} K_{ij},$$

and their overall sum $K_{iS} = K_{iA} + K_{iB}$. Under the null, we can compute the probabilities of the events $K_{iA} = a$ and $K_{iB} = b$ for any pair of numbers a and b. They denote this probability by $p(a, b)$. Then the sum of all probabilities less or equal to $p(k_{iA}, k_{iB})$, given that the overall sum is k_{iS}:

$$p_i = \frac{\sum_{a+b=k_{iS}\ p(a,b)<p(k_{iA},k_{iB})} p(a,b)}{\sum_{a+b=k_{iS}} p(a,b)}.$$

The variables a and b in the above sums take the values $0, \ldots, k_{iS}$.

3.2.3 NBPSeq

NBPSeq uses an over-parameterized version of the negative binomial called the NBP

distribution, which incorporates a non-constant dispersion parameter directly within a parametric family. To make clear the difference between NBPSeq and DESeq, let′s recall the variance modeling part in the following form:

$$DESeq: \sigma_{ij}^2 = \mu_{ij}(1 + \phi_\mu \mu_{ij})$$

$$NBPSeq: \sigma_{ij}^2 = \mu_{ij}(1 + \phi \mu_{ij}^{\alpha-1})$$

Let $NBP(\mu, \phi, \alpha)$ denote a NBP distribution with mean μ and dispersion parameters (ϕ, α). The probability mass function of a $NBP(\mu, \phi, \alpha)$ random variable Y can be expressed as follows:

$$Pr(Y = y; \mu, \phi, \alpha) = \frac{\Gamma(\gamma + y)}{\Gamma(\gamma)\Gamma(1 + y)}(1 - p)^\gamma p^y,$$

where $p = \frac{\mu}{\mu+\gamma}$ and $\gamma = \phi^{-1}\mu^{2-\alpha}$, for $y = 0,1,2, \ldots$.

Once relative frequencies π_{ik} is estimated using the method of moments,

$$\widehat{\pi_{ik}} = \frac{1}{mJ}\sum_{j=1}^{J} y_{ijk},$$

then the remaining processes is exactly the same as DESeq. Specifically, performing a statistical test for the null hypothesis

$$H_0: \pi_{i1} = \pi_{i2}$$

for each gene i based on the statistical evidence provided by the sequence read counts Y_{ijk} in two treatment groups $k = 1,2$. The two-sided p-value can be computed using exactly the same procedure of DESeq (Anders and Huber 2010).

3.2.4 NOISeq

NOISeq is a nonparametric approach. NOISeq takes corrected value or pseudo-counts x_{gj}^{i} to obtain the statistics needed to derive DE. The DE statistics in NOISeq are the log-ratio (M) and the absolute value of difference (D), defined for a gene i as $M^i = \log_2 \frac{x_1^i}{x_2^i}$ and $D^i = |x_1^i - x_2^i|$.

Let M* and D* be the random variables describing noise distribution, which is drawn from replicates in the same experimental condition. Let G^i be a random variable that takes the value 1 if gene i is differentially expressed between conditions, and takes 0 otherwise. Then, the probability of a gene being DE given the expression levels in both conditions can be written as follows:

$$P\left(G^i = 1 \middle| x_1^i, x_2^i\right) = P\left(G^i = 1 \middle| M^i = m^i, D^i = d^i\right) = P(|M^*| < \left|m^i\right|, D^* < d^i)$$

Thus the probability of not being DE between the two conditions can be easily derived as $P\left(G^i = 0 \middle| M^i = m^i, D^i = d^i\right) = 1 - P(|M^*| < \left|m^i\right|, D^* < d^i)$. The odds $\frac{P\left(G^i = 1 \middle| M^i = m^i, D^i = d^i\right)}{P\left(G^i = 0 \middle| M^i = m^i, D^i = d^i\right)}$ can be used decide whether a gene is DE or not. NOISeq recommends threshold of 0.8, which is equivalent to odds value of 4:1.

3.3 Local Pooled Error (LPE) Method

In this section, a method, named "Local Pooled Error with RNA-Seq" (LPEseq), is proposed for the analysis of DE detection with a small number of replicates. Based on the local pooled error (LPE) method in microarray analysis (Jain, Thatte et al. 2003), additional steps are included to allow the method to be applicable to RNA-Seq data with a non-replicated set. The method assumes that the two singleton experiments from different conditions are replicates and the outliers (possibly originated by a mixture of two different distributions) are measurement errors and removed. On the basis of this additional steps, it is possible to evaluate of local pooled error from different conditions with less affected by outliers. The method is substantially superior to the others, especially when no replicates are available. In section 3.3.1 provides a brief introduction of LPE method, and in 3.3.2, LPE method will be re-visited in depth. Section 3.3.3-4 will describe how I extend the original method with a toy example. Section 3.3.5 makes a comparison with other existing methods.

3.3.1 A Brief Introduction to LPE method

Original idea of LPE method is presented here. The original LPE method, which pools the error in each local intensity bin and shrinks each error variance estimate toward the mean of other probes (or genes) with similar intensities, was developed for microarray experiments in which gene expression intensity is continuous values

(Jain, Thatte et al. 2003). In brief, the method first evaluates the baseline error distribution for each of the compared experimental conditions, say class $\mathbb{X}$ and $\mathbb{Y}$, respectively. For duplicated arrays (subscript with 1 and 2), for instance, the variance of $M\ (= x_1 - x_2\ \&\ x_2 - x_1)$ on predetermined quantiles of $A\ (= (x_1 + x_2)/2)$ is evaluated and then a cubic smoothing spline is fit to the variance estimates on the quantiles. The baseline error distribution for condition Y is derived in the same way. Test statistic of LPE method is calculated as follows:

$$Z = \frac{Med(X) - Med(Y)}{\sigma_{pool}}$$

where

$$\sigma^2_{\text{pool}} = k\left(\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right),$$

n_X and n_Y are the number of replicates in the two array samples being compared; σ_X^2 and σ_Y^2 are the estimates of variance of the transcript in condition $\mathbb{X}$ and $\mathbb{Y}$; k is a scaling factor for using median (Jain, Thatte et al. 2003).

The LPE method assumes that genes with similar observed intensities have similar expression variances. Based on this assumption, the LPE method estimates gene specific variance from a calibration curve derived from pooling the variance estimates of replicated expression differences of genes within similar expression intensities. Since the LPE method is based on calculating error variance from replicated experiments in the same condition, it cannot be directly applied to

experiments with no biological replicates in one or both of the conditions. Furthermore, the LPE method was developed for continues variables, unlike the read counts measured in RNA-Seq. Thus a new method should address these issues.

3.3.2 LPE method revisited

This section describes details of LPE method (with replicate in each condition) to fully appreciate the method and to extend it to none-replicated RNA-Seq experiments. Index i represents a transcript or a gene, j indicates a replicate (either of biological or technical), (k) represents a k_{th} bin, and l_k represents the transcripts in a bin (k). With these indexes, X_{ij} and Y_{ij} are log transformed intensity of a transcript i obtained from the replicates j in condition $\mathbb{X}$ and $\mathbb{Y}$.

Here LPE makes three following assumptions. (i) $X_{ij} \sim \left(\mu_{i,X}, \sigma_{i,X}^2\right) \leftrightarrow X_{ij} = \mu_{i,X} + U_{ij}$ where $\mu_{i,X}$ is fixed value and U_{ij} is random variable following $\left(0, \sigma_{i,X}^2\right)$, (ii) $\forall l_k$ where $l_k \in k_{th}$ bin, $M_{(k)l_k}$, defined by the difference between observations, following $(0, \sigma_{(k)}^2)$.

The first step is a variance evaluation step. Here, a case with duplicated arrays in each condition is explained for simplicity. Generalization to more than duplicated arrays $(j = 1, 2)$ are straightforward. To evaluate the baseline error distribution of $\mathbb{X}$, mean intensity of all transcripts $A_{i,X}\left(= \frac{X_{i1} + X_{i2}}{2}, i = 1, \ldots, \mathrm{N}\right)$ are evaluated and then quantile bins (percentiles by default) are constructed using theses $A_{i,X}$ values. For in a bin (k), the variance of $M_{(k)l_k,X}(= X_{l_k 1} - X_{l_k 2})$ on pre-determined

quantiles are evaluated. Note that with more than duplicates, all pairwise comparisons of (M, A) are used.

Here original LPE method used both $M_{(k)l_k,X}$ and $-M_{(k)l_k,X}(= X_{l_k2} - X_{l_k1})$ values to estimate pooled variance $\sigma^2_{(k)}$ in a bin (k). However this might affect the estimation performance and need to be checked. The problem can be simplified as follow. Let $M_1, \dots, M_n$ are random sample from $(0, \sigma^2_{(k)})$, then $\bar{M} = \frac{1}{n}\sum_i^n M_i$ and $s^2 = \frac{1}{n-1}\sum_i^n (M_i - \bar{M})^2$. Denote sample mean and sample variance of 2n of samples $M_1, \dots, M_n, -M_1, \dots, -M_n$ as $\bar{M}_{2n}$ and s^2_{2n}. By symmetry, $\bar{M}_{2n} = 0$ and after some calculus, one can easily obtain $s^2_{2n} = \frac{2n-2}{2n-1}s^2 + \frac{2n}{2n-1}\bar{M}$. This equation can be more rigorously written as $\hat{\sigma}^2_{(k),2n} = \frac{2n-2}{2n-1}\hat{\sigma}^2_{(k),n} + \frac{2n}{2n-1}\bar{M}_{(k),n}$. Here $\hat{\sigma}^2_{(k),2n}$ is biased but consistent estimator.

Biasness:

$$\mathrm{E}\left[\hat{\sigma}^2_{(k),2n}\right] = \frac{2n-2}{2n-1}\mathrm{E}\left[\hat{\sigma}^2_{(k),n}\right] + \frac{2n}{2n-1}E\left[\bar{M}_{(k),n}\right] = \frac{2n-2}{2n-1}\sigma^2_{(k)} + 0 \neq \sigma^2_{(k)}$$

Consistency:

$$\operatorname*{plim}_{n\to\infty} \hat{\sigma}^2_{(k),2n} = \operatorname*{plim}_{n\to\infty} \frac{2n-2}{2n-1}\hat{\sigma}^2_{(k),n} + \operatorname*{plim}_{n\to\infty} \frac{2n}{2n-1}\bar{M}_{(k),n} = \sigma^2_{(k)} + 0$$

<< lemma>>

$\operatorname*{plim}_{n\to\infty} \hat{\sigma}^2_{(k),n} =^? \sigma^2_{(k)}$

<proof>

$$\hat{\sigma}^2_{(k),n} = \frac{1}{n-1}\sum_{i=1}^{n}(M_i - \bar{M})^2 = \frac{n}{n-1}\left\{\frac{1}{n}\sum M_i^2 - \left(\frac{1}{n}\sum M_i\right)^2\right\}$$

$$\operatorname*{plim}_{n\to\infty} \hat{\sigma}^2_{(k),n} = \operatorname*{plim}_{n\to\infty} \frac{n}{n-1}\left\{\frac{1}{n}\sum M_i^2 - \left(\frac{1}{n}\sum M_i\right)^2\right\} = 1 \times \mathrm{E}(M_1^2) - [E(M_1)^2] = \sigma^2_{(k)}$$

I performed the comparison between $\hat{\sigma}^2_{(k),n}$ and $\hat{\sigma}^2_{(k),2n}$ by simulation study.

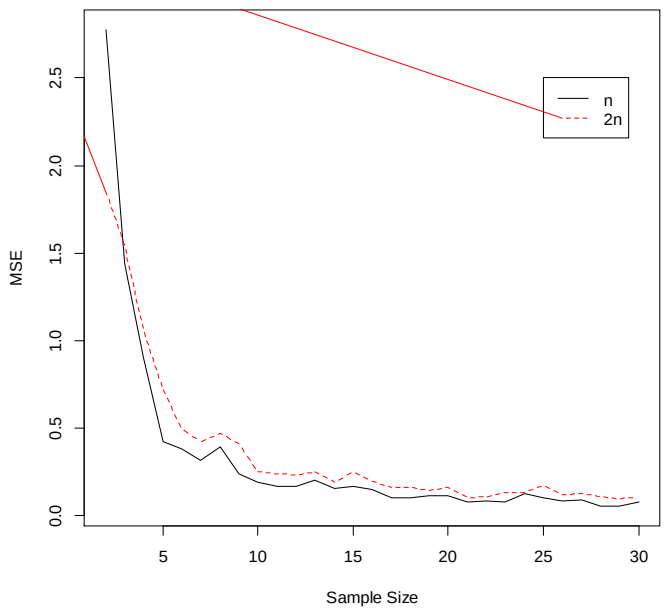


Figure 3-2 Comparison of estimator performance. MSE of each estimator is plotted.

i) First generate random sample of size n from N(0, 1).

ii) Then evaluate sample variance using n observations and 2n observation using $\hat{\sigma}^2_{(k),2n} = \frac{2n-2}{2n-1}\hat{\sigma}^2_{(k),n} + \frac{2n}{2n-1}\bar{M}_{(k),n}$.

iii) Repeat the step with varying n (2 to 500)

iv) For each n, repeat 100 times and evaluate $\frac{1}{k}\sum_{k=1}^{100}\left(\hat{\sigma}^2_{(k),2n}-1\right)^2$ and $\frac{1}{k}\sum_{k=1}^{100}\left(\hat{\sigma}^2_{(k)}-1\right)^2$.

v) Plot the results

For both estimator well performed for large sample size (Figure 3-2). However, for extremely small sample sizes (n=2,3), $\hat{\sigma}^2_{(k),2n}$ shows smaller MSE value than $\hat{\sigma}^2_{(k),n}$. This is because, for small sample, $\hat{\sigma}^2_{(k),n}$ can be underestimated when observations are same-signed. Recall that these samples are from symmetric around mean zero. The following simple example clarifies the result more clearly. Consider three observations, 0.9, 0.5, 0.4 from N(0, 1), then $\hat{\sigma}^2_{(k),n}=0.07$, far smaller than true value 1. However, $\hat{\sigma}^2_{(k),2n}\sim 0.5$, much improved by adding mean of three observations.

Once the variances of M's are calculated, then a smooth local regression curve is fit to the variance estimates on quantiles in each condition $\mathbb{X}$ and $\mathbb{Y}$ (Figure 3-3). Based on these curves, the variance of each transcript can be estimated by plugging

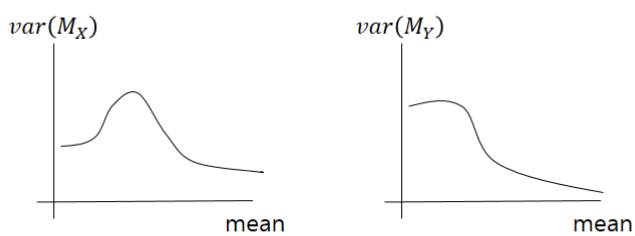


Figure 3-3 Baseline error distribution for $\mathbb{X}$ and $\mathbb{Y}$.

the mean intensity of the transcript into the curve. Note that $X_{ij} \sim (\mu_{i,X}, \sigma_{i,X}^2) \leftrightarrow X_{ij} = \mu_{i,X} + U_{ij}$ where $\mu_{i,X}$ is fixed value and U_{ij} is random variable following $(0, \sigma_{i,X}^2)$ and the baseline error curve is evaluated using $M_{(k),X}$ $(= X_{(k)l_k1} - X_{(k)2})$. Thus $var(M_{(k),X}) = var(U_{(k)1} - U_{(k)2}) = var(U_{(k)1}) + var(U_{(k)2}) = 2 \times \sigma_{i,X}^2$.

For transcript i, differential expression test can be performed using test statistic $z_i = \frac{\tilde{X}_i - \tilde{Y}_i}{\hat{\sigma}_{i,pooled}}$, where $\tilde{X}_i$ and $\tilde{Y}_i$ are median value of transcript i in condition $\mathbb{X}$ and $\mathbb{Y}$, respectively and $\hat{\sigma}_{i,\text{pooled}}^2 = \frac{(n_1-1)\hat{\sigma}_{i,X}^2 + (n_2-1)\hat{\sigma}_{i,Y}^2}{n_X + n_Y - 2}$ with n_X and n_Y are the number of replicates or samples in $\mathbb{X}$ and $\mathbb{Y}$.

3.3.3 Extension to LPEseq

This section relates the two ideas, reviews LPE and its properties in more detail, and improves the method by focusing on two aspects of refinement – applicability to RNA-Seq and the experiments with no replicates in each condition. For the former I will describe how I addressed the count values into the method. For the latter, I treated two non-replicated samples as if they were replicates and removed the outliers so to make a less impact on LPE estimation. Thus, the refinement of LPE method retains the original concept (Figure 3-4).

To be more specific, let's focus on the problem of inferring DE between two different conditions. Let again X'_{ij} and Y'_{ij} represent the number of read counts mapped to a specific transcript (or gene) i in the j_{th} sample (or replicate or lane) from the experimental condition or class $\mathbb{X}$ and $\mathbb{Y}$, respectively. Since X'_{ij} and Y'_{ij} are affected by the depth of sequencing, those values are not directly comparable. The relative abundance of transcripts across the samples should be normalized. By dividing the count of each transcript by the total number of read counts for that experiment, the count values from different experiments become comparable. Then

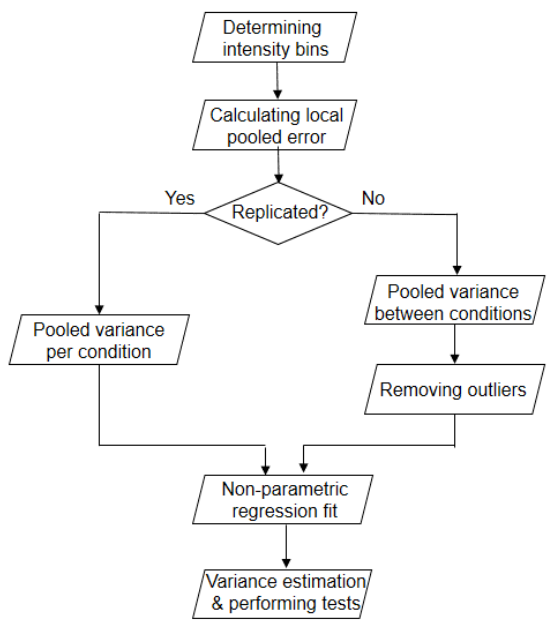


Figure 3-4 A schematic representation of the proposed method. The flow chart of the proposed algorithm is shown.

normalized log-transformed intensity values, $X_{\mathrm{ij}} = \log_2 \frac{X'_{\mathrm{ij}}}{\sum_i X'_{ij}}$ and $Y_{ij} = \log_2 \frac{Y'_{\mathrm{ij}}}{\sum_i Y'_{ij}}$ are calculated. Once these values are obtained, the LPE can be derived by the same manner in as described in (Jain, Thatte et al. 2003).

Without replicates, only single measurement exists in each class. Let X_i and Y_i denote the normalized log-transformed count values of transcript i under two different conditions $\mathbb{X}$ and $\mathbb{Y}$, respectively. Note that without replicates ($j = 1$), j index can be dropped without ambiguity. Let $X_i \sim (\mu_{\mathrm{i,X}}, \sigma^2_{i,X})$ and $Y_i \sim (\mu_{\mathrm{i,Y}}, \sigma^2_{i,Y})$, where $\mu_{\mathrm{i,X}}$, $\sigma^2_{i,X}$ are mean and variance of transcript i in condition $\mathbb{X}$. Similar for $\mathbb{Y}$. If two random variables, $U_i (\sim (0, \sigma^2_{i,X})$ and $V_i (\sim (0, \sigma^2_{i,Y}))$ are introduced, X_i and Y_i can be re-written as $X_i = \mu_{i,X} + U_i$ and $Y_i = \mu_{i,Y} + V_i$. Under the null (H_0: $\mu_{i,X} = \mu_{i,Y}$), LPEseq assumes that $\sigma^2_{i,X} = \sigma^2_{i,Y} = \sigma^2_{\mathrm{i}}$.

Since no replicates are available, the baseline error distribution is obtained by regarding each sample in different conditions as replicates. Again evaluating qunatile bins is the first step of LPEseq method. A_i $(= \frac{X_i + Y_i}{2})$ values are evaluated with two observations each from different condition $\mathbb{X}$ and $\mathbb{Y}$ and quantile bins are constructed. Then the variance of M $(= X_{l_k} - Y_{l_k})$ on pre-determined quantiles of

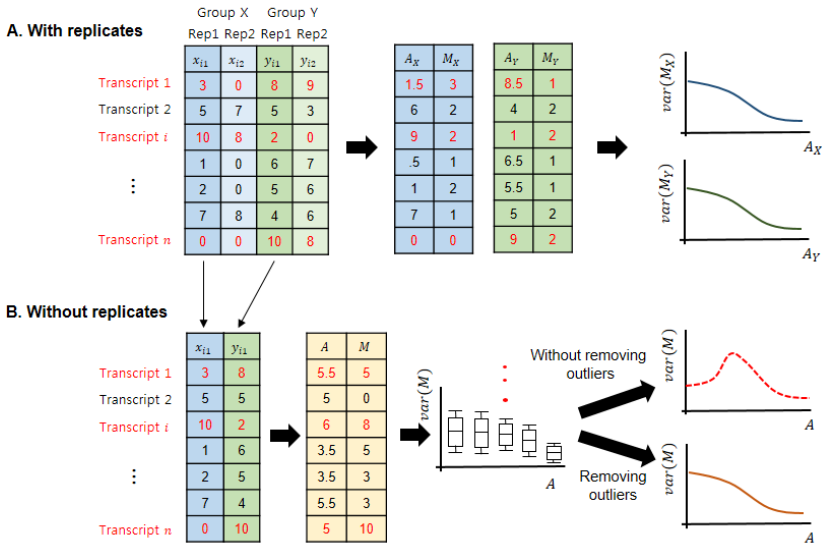


Figure 3-5 A schematic diagram of the proposed method. Local pooled error is evaluated from the replicates of each class (A). When replicates are not available, however, local pooled error can be evaluated by regarding two singleton experiments in different classes as replicates (call it 'pseudo replicate'). Under this 'pseudo replicate' assumption, outliers (denoted in red) can lead unexpected overestimation of local pooled error. Therefore an auxiliary step, removing outliers, is included when working without replicates (B).

A is calculated. In this case, however, since variance of M is not drawn from the same condition, differentially abundant transcripts can adversely affect the proper evaluation of the LPE per each condition.

3.3.4 A toy example

To better understand the problem described in the previous section, suppose there

have 3 DE transcripts among n transcripts as shown in Figure 3-5 (DE transcripts are coloured in red in the figure). With replicate information, the LPEs are drawn from each conditions separately (Figure 3-5A), whereas those from non-replicated conditions have to be evaluated from single observations of different conditions (Figure 3-5B). When evaluated with data with no-replicates per each condition, the LPEs of three DE transcripts (coloured in red in the Figure 3-5) show clear distinction compared to those with replicates per each condition and thus, a fitted variance curve can be biased (a red dashed curve in Figure 3-5B). By calculating outlier scores of given data in each quantile, one can remove extreme observations (or outliers) and re-calibrate a local pooled error curve less affected by these outliers (an orange solid curve in Figure 3-3B). For scoring outliers, any kind of outlier detection methods can be applied, including Z-score ($Z_i = \frac{x_i - \bar{x}}{\mathrm{sd}}$, where $X_i \sim N(\mu, \sigma^2)$, and sd denotes the standard deviation of data), MAD score ($MAD = median(|x_i - \tilde{x}|)$ where $\tilde{x}$ is the sample median), modified Z-score ($Z_i^* = \frac{x_i - \bar{x}}{MAD}$), and so on.

We summarize the procedure of LPE analysis without replicate as follow.

Determining intensity bins

I. Calculate the mean intensity of transcript i in two different conditions (*i.e.*, $(x_i + y_i)/2$).

II. Calculate quantiles (percentiles by default) with mean intensities of whole transcripts, evaluated at step I and define 'intensity bins' using adjacent quantile values. Then, place the transcripts in the bin where their mean intensities belong.

Detecting outliers in a bin

III. Evaluate the modified Z-scores of all transcripts i $(i: 1 \dots, l_k)$ in each bin using the following equation,

$$\chi^2 = \frac{\left(M_{(k)l_k} - \overline{M}_{(k)}\right)^2}{var(M_{(k)})}$$

IV. Label the transcripts as outliers when p-value ≥ 0.95

Evaluating local pooled error without outliers

V. Remove outliers labelled in step IV, then evaluate variance in each bin.

VI. Generate local pooled error curve by fitting a cubic smoothing spline to the variance along with the bins. This makes the LPE as a function of transcript abundance level.

VII. Use the LPE function drawn in step VI to estimate the transcript-specific variance by plugging the value of each transcript.

Once the variance in each experimental condition is derived, testing a hypothesis on differential expression is similarly conducted as is done in the analysis with

Effects of DE transcripts (A numerical example – DE case)

	X	Y	A	M
1	0.5	0.5	0.5	0
2	0	1.2	0.6	-1.2
3	1	0.1	0.55	0.9
4	0	1.8	0.9	-1.8
5	1.5	2.3	1.9	-0.8
6	3	0.8	1.9	2.2
7	2.8	0.5	1.65	2.3
8	0.7	0.9	0.8	-0.2
9	1.8	2	1.9	-0.2
10	1.8	1.5	1.65	0.3

(step 1) calculating quantiles using A

0	25%	50%	75%	100%
0.5	0.65	1.275	1.8375	1.9

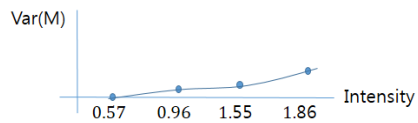
(step 2) Place transcripts into the bins using A and calculate varM

Bin	$bin_{(1)}$ [0.5,0.65]	$bin_{(2)}$ [0.65, 1.275]	$bin_{(3)}$ [1.275,1.8375]	$bin_{(4)}$ [1.8375, 1.9]
genes	1, 2, 3	4, 8	7,10	5, 6, 9
M	0, -1.2, 0.9 0,1.2,-0.9	-1.8, -0.2 1.8, 0.2	2.3, 0.3 -2.3, -0.3	-0.8, 2.2, -0.2 0.8, -2.2, 0.2
Var(M)	0.9	2.186	3.586	2.84

(additional step) Removing outliers $\chi^2_{(k)} = \frac{\left(M_{(k)l_k} - \bar{M}_{(k)l_k}\right)^2}{var(M_{(k)l_k})}$

Var(M)	0.0	0.08	0.18	0.453

(step 3) baseline error distribution curve using center of each bin and var(M) without outliers



(step 4) DE test

$$z = \frac{\bar{x}_i - \bar{y}_i}{\hat{\sigma}_p} \sim N(0,1), \quad \hat{\sigma}_p^2 = \frac{1}{2}\left(\hat{\sigma}_{i,X}^2 + \hat{\sigma}_{i,Y}^2\right)$$

Both $\hat{\sigma}_{i,X}^2$ and $\hat{\sigma}_{i,Y}^2$ are from $\hat{\sigma}_i^2 = \frac{1}{2}var(M)$

Figure 3-6 A toy example with numerical values. An artificial toy example is presented with ten transcripts. Five red-coloured transcripts denote the transcripts with differential expression.

replicates. Figure 3-6 shows a toy example with artificial numeric values of ten transcripts.

3.3.5 Comparison with other methods

LPEseq is compared with edgeR (Robinson and Oshlack 2010), DESeq (Anders and Huber 2010), NBPSeq (Di, Schafer et al. 2011) and NOISeq (Tarazona, Garcia-Alcalde et al. 2011) with respect to the power and false discovery rate. All transcripts found to be DE at a FDR threshold of 0.05 was considered significantly DE. It is not clear, however, how to set a threshold for the threshold returned by NOISeq to be comparable with the FDR estimate from the other methods. Therefore, to include NOISeq in comparison, we set up the threshold value for NOISeq to tolerate about 5% of false positives from null simulated datasets. We have mostly used the default setting provided in the review article, but for the non-replicated data analysis we used recommended options provided in the implementations.

3.3.6 Additional Extension of the LPEseq method

The key problem in extending LPE method to none-replicated experiments lays in estimating variance. The key idea of handling this issue, made in section 3.3.2, was to use two observations, each made from different conditions, as 'pseudo replicates' and to evaluate the local pooled variance using them. Here, local means that the transcripts with similar intensity.

When defining local transcripts, transcripts were grouped using their mean intensity A, calculated using two observations in different conditions. However, different values can be used for the same purpose. For example, let's suppose we

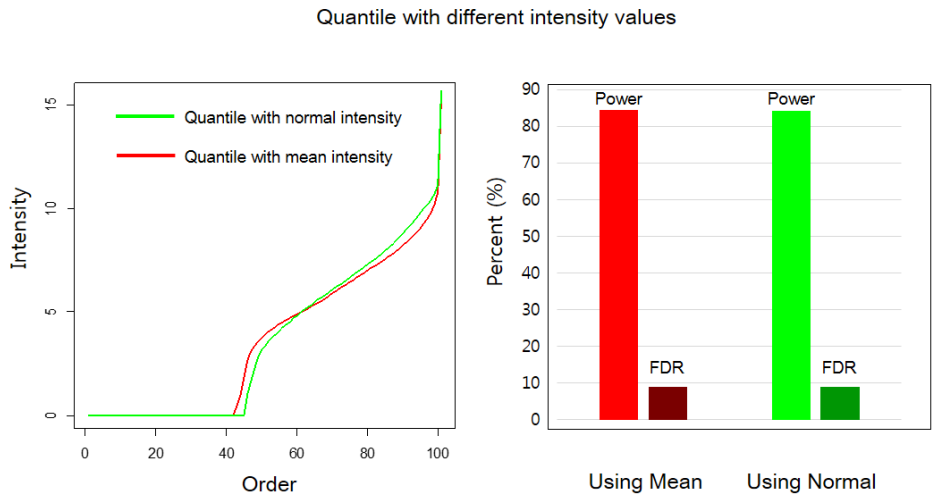


Figure 3-7 Comparison of quantiling processes with different intensity values. Red and green colour indicate the results made using mean and normal intensity values, respectively. Quantiles evaluated using mean and normal intensity values are shown (left). In simulation performance (DE portion: 10%, count difference of DE: 1000, dispersion: 0.27), with quantiling process using mean and normal values (right)

perform a DE analysis between normal and cancer sample. Since we do not have replicated samples in each class, we might regard them as replicates and evaluate local pooled error based on these. In calculating LPE step in previous section, the proposed method uses A value of two samples. But one might worry about using average intensity, because averaging two values from different conditions might distinct from averaging with same conditions. Therefore the quantiles evaluated using A intensities and using intensities in one of two classes are compared. As shown in Figure 3-7, no significant difference was observed between two processes (Figure 3-7). Performance were evaluated with power and false discovery rate (FDR)

using simulation datasets. What observed was not only the similar behavior in quantile plot (Figure 3-7 left), but also their power and FDR in various simulation settings (only one case is depicted in Figure 3-7 right).

3.4 Real Data Application

3.4.1 Preparing real datasets

In this study, we present results based on the following four real RNA-Seq datasets, which had been pre-processed and distributed by ReCount (Frazee, Langmead et al. 2011). We selected four datasets to analyze how well some characteristics of our proposed method corresponds to those of other competing methods in various situations: These datasets with different replication types, conditions (different classes or same classes) and numbers (from 1 to 30 per each class) were analyzed to see how well some characteristics of our proposed method corresponds to those of other competing methods. We summarized the characteristics of each dataset below.

A. Sultan dataset (two replicates in each condition): Sultan et al. (Sultan, Schulz et al. 2008) performed RNA-Seq experiments in a human embryonic kidney and a B cell line. In each sample of this data, two replicates were generated, and we analysed differential expression between two conditions. The dataset provides a case in which a small number of replicates are available.

B. Hammer dataset (single replicate in each condition): Hammer et al. (Hammer, Banck et al. 2010) studied rats with chronic neuropathic pain induced by spinal nerve ligation (SNL) in serial experiments using RNA-Seq. Rats with SNL for 2 weeks and 2 months were sequenced and compared with controls. We performed DE analysis between SNL and control for 2 weeks. The dataset provides a case in

which only single replicate are available in each condition.

C. HapMap dataset (60 replicates in the same condition): As a part of the International HapMap Project (International HapMap 2003), Montgomery et al. performed RNA-Seq in human lymphoblastoid cell lines derived from 60 CEU samples (Montgomery, Sammeth et al. 2010). As each tissue culture was derived from a different subject and so has a different genotype, these data show high variability. This dataset provides a case of null data (under the same condition) with a large number of replicates.

D. MAQC dataset (seven replicates and one pooled sample in each condition): Two types of samples, Ambion's human brain reference RNA and Stratagene's human universal reference RNA, were assayed by Bullard et al. (Bullard, Purdom et al. 2010), and herein referred to as Brain and UHR, respectively. Both seven technical replicates and pooled count from the replicates are available for this dataset in the ReCount database (Frazee, Langmead et al. 2011). This dataset provides not only a case with large replicates, also a case for a comparison between results with replicates against without replicates.

3.4.2 Results

We have applied LPEseq to analyze four real datasets, each of which has distinct characteristics: non-replicates (Hammer dataset), small replicates (Sultan dataset), large replicates under the null (HapMap dataset), and both replicates and non-

replicates (MAQC dataset). The purpose of this analysis is to discover DE transcripts and compare the results found by other competing methods. Among several methods recently developed, we mainly compare LPEseq with edge.R, DESeq, NBPSeq and NOISeq with respect to DE detection ability and overlaps among them. The reasons for choosing these methods are (i) that a recent review performed many methods and recommended edgeR, DESeq and NBPSeq as powerful tools for testing DE with the smallest sample size and (ii) that NOISeq is the only one of all methods which can be applicable to non-replicates, except DESeq.

First, we compared the number of DE transcripts found by each method (Figure 3-8). For the Sultan dataset, which has small number of replicates, a similar number of DE transcripts were found by all methods excluding NOISeq. NOISeq called about two times more DE transcripts than others (Figure 3-8A). Without replicates, only LPEseq, DESeq and NOISeq are applicable, thus we compared these three methods using the Hammer dataset. Hammer *et al.* reported about 2000 transcripts were DE and validated them with qPCR (Hammer, Banck et al. 2010). Interestingly, LPEseq obtained 1995 DE transcripts, a similar number of DE transcripts that were reported by the Hammer *et al.*, while NOISeq and DESeq detected 6099 and 74 transcripts as DE, respectively (Figure 3-8B). For the MAQC dataset with replicates, edgeR, DESeq, NBPSeq, and NOISeq detected approximately the same number of DE transcripts, while LPEseq returned relatively a half of them (Figure 3-8D and Figure 3-9).

For the HapMap dataset, consisting of only normal samples, we expect that no

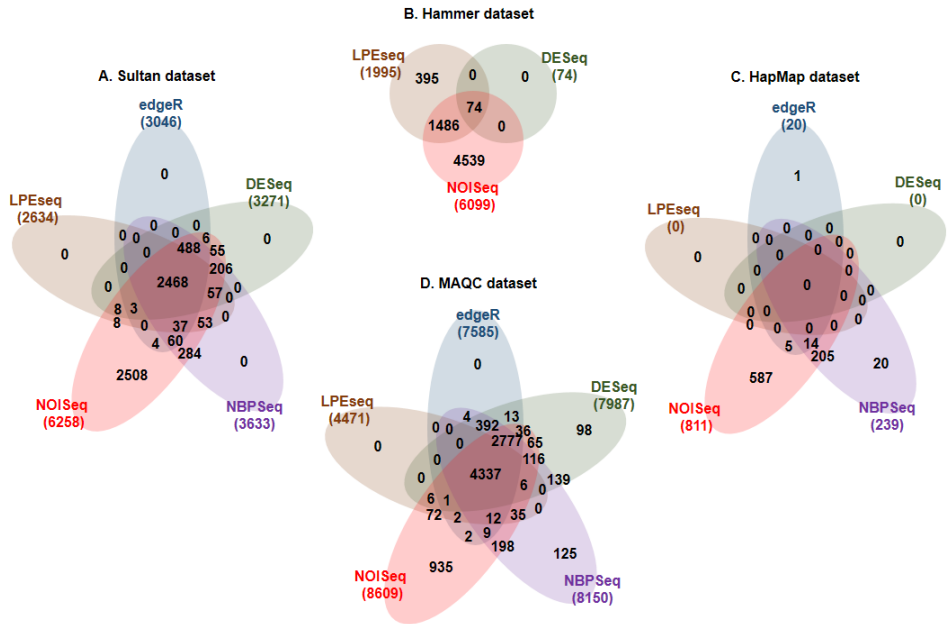


Figure 3-8 Venn diagrams of DE transcripts detected by five methods. Three different datasets, a small number of replicates (A) or non-replicates (B) of different conditions and a large number of replicates in the same condition (C), and a large number of replicates in two classes (D) were used to compare the number of DE transcripts and the overlaps detected by LPEseq (brown), edgeR (sky blue), DESeq (green), NBPSeq (violet), and NOISeq (red). The number in parentheses indicates the total number of DE transcripts found (DE calls with FDR < 0.05 criteria for all methods; please see 'Materials and Methods' section for further details)

transcripts are truly DE. Nevertheless, only LPEseq and DESeq detected no DE transcripts (Figure 3-8C). NOISeq found 811, the largest number of DE transcripts, which is similar to previous observations with other datasets. NBPSeq and edgeR found 239 and 20 transcripts were significant DE, respectively (Figure 3-8C).

Next, we studied the overlap between the sets of transcript called DE by all

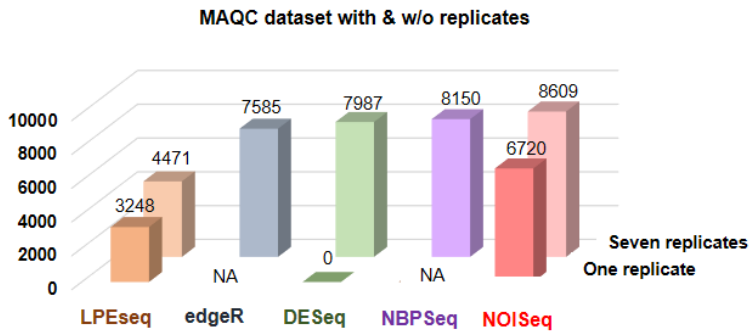


Figure 3-9 A bar plot representing the effect of the existence of replicates on detecting DE transcripts. Colours and brightness denote the method applied and the number of replicates used, respectively. Here (in order), brown, sky blue, green, violet, and red represent LPEseq, edgeR, DESeq, NBPSeq, and NOISeq. The brighter colour indicates the results without replicates.

methods. The DE transcripts found by LPEseq were to a large extent found also by other methods in many cases. Over 90% of DE transcripts found by LPEseq were also by the other four methods in the Sultan dataset and the MAQC dataset with replicates (Figure 3-8A and 3-8D). Also a large portion (~78%) of DE transcripts found by LPEseq was shared by NOISeq (Figure 3-8B). In this dataset, however, DESeq detected few transcripts.

To further evaluate the consistent performance of the methods with and without replicates, we applied them to the MAQC dataset without replicate, consisting of only one pooled sample per each condition. By comparing to the results obtained by

the same dataset with replicates, we expect that a consistent number of DE transcripts can be found. As can be seen in Figure 3-6, both LPEseq and NOIseq showed relatively a small decrease (1223 and 2089) in DE detection, while DESeq detected none from 7987 DE transcripts (Figure 3-9).

By studying the transcripts called DE in real data analyses, we noted that the most of DE transcripts found by LPEseq overlapped with those called by other methods. Among the overlaps, DE transcripts found by NOISeq often overlapped more with those called by NBPSeq, while transcripts found by edgeR overlapped more with those from DESeq. The highest number of DE transcript was found by NOISeq in all cases, even with the null case dataset.

3.5 Simulation Study

The genome-wide association study (GWAS) is an approach that involves rapidly scanning single-nucleotide polymorphism (SNP) markers across genome

3.5.1 Generating simulation datasets

The purpose of the simulation study is to investigate the ability of DE detection under varying effects of 4 different factors: the effect size (counts difference) between conditions, the portion (number) of DE transcripts, dispersion parameter in NB distribution, and the number of replicates in each condition. To generate the simulation datasets, we adopted the assumption which was made in many other published works [ref], i.e., the abundance of transcript i, denoted by X_i follows $NB(\mu_i, \phi_i)$ and $\mu_i = d + \nu_i + \delta_i I_{(i \in DE)}$. Here, d captures the sequencing depth of experiment, ν_i represents specific abundance of transcript i in the one of the class, δ_i is the differential expression counts, and $I_{i \in DE}$ indicates whether transcript i is DE of not. Note that under the sample experiment, μ_i only depends on ν_i.

We estimated (μ_i, ϕ_i) using the Montgomery *et al.*'s real data set, one of the largest sample datasets, including 60 unrelated normal Caucasian individuals. The log-likelihood function for N iid variables from a Negative Binomial distribution, given counts $x_{i1}, \ldots, x_{iN}$, is

$$\ell(\mu_i, \kappa_i | x_{i1}, \dots, x_{iN})$$

$$= \sum_{j=1}^{N} \log\Gamma\left(x_{ij} + \frac{1}{\phi_i}\right) - N\log\Gamma\left(\frac{1}{\phi_i}\right) - \sum_{j=1}^{N} \log\Gamma\left(x_{ij} + 1\right)$$

$$- \sum_{i=1}^{N} x_{ij} \log\left(\frac{\mu_i \phi_i}{1 + \mu_i \phi_i}\right) - \frac{N}{\phi_i} \log(1 + \mu_i \phi_i).$$

The maximum likelihood estimate (MLE) of μ_i was first obtained for each transcript across 60 samples, then ϕ_i was estimated by maximizing the log-likelihood function numerically using the Simulated ANNealing algorithm (Henningsen and Toomet 2011).

In this work, we focused our attention to DE analysis under the two different classes with and without replicated samples. Briefly it is as follows: i) μ_i for all transcripts ($i = 1, \dots, 20000$) were randomly selected from the estimated values, ii) for each transcript i, we generated transcript counts, x_{ij} and y_{ij} from $NB(\mu_i + \delta I_{i \in DE}, \phi_i = \phi)$ in each class and $k\%$ of the transcripts were set to be differentially expressed with effect size δ. In our simulation, we analyzed 360 different combinatorial situations according to the values of k, δ, ϕ and the number of replicates as stated below.

Case I: Different effect size: (100, 500, 1000, 5000, 10000 and 15000)

- We varied the effect size ranging from 100 to 15000, which is roughly the range of mean and maximum count difference between two classes in real dataset with similar total read counts of simulated data.

Case II: Different portion of DE transcripts: (1%, 10% and 20%)

- The number of DE transcripts can varies due to the biological phenomena of interest and can affect the DE detection performance of the tests. We set 1% - 20% of transcripts as to be DE and observed their effects on the DE detection ability.

Case III: Different dispersion: (0.01, 0.89, 0.2, 0.27 and 0.41)

- We use interquantile range [0.2, 0.41] of the estimated values for the dispersion parameter ϕ. We also include 0.01 (Poisson-like behaviour) and 0.89 (minimum of the estimates) to observe the performance of the method with small dispersion.

Case IV. Different number of replicates per each class: (1, 2, 3, and 10 replicates)

- We repeated the analyses of case I – III with the different number of replicates per each class, focusing on small sample analysis.

3.5.2 Simulation results

We performed intensive simulation studies to investigate the performance of LPEseq in aspect of True DE detection ahead of false discovery under a variety of situations.

Four different parameters were investigated and the details are described in Materials and Methods. In most of our simulation settings we observed similar patterns, hence, we summarize a part of results obtained from following cases:

- Case 1: Different effect sizes (with 0.27 dispersion, 10% DE, 2 replicates)
- Case 2: Different dispersion (with 1000 effect size, 10 DE, 2 replicates)
- Case 3: Different DE portion (with 1000 effect size, 0.27 dispersion, 2 replicates)
- Case 4: Different number of replicates (with 1000 effect size, 0.27 dispersion, 10% DE)

We evaluated the ability to detect true DE transcripts above non-DE transcripts in all cases (Figure 3-10). Specifically, we observed power ($\frac{\text{\# of true DE transcripts detected}}{\text{\# of true DE transcripts}}$) and false discovery rate ($\frac{\text{\# of false DE transcripts detected}}{\text{\# of DE transcripts detected}}$). In all cases, NOISeq was the most powerful method but at the price of a remarkably higher false discovery rate, which supports our previous observations made with real datasets. Therefore, unless otherwise stated, NOISeq is out of description.

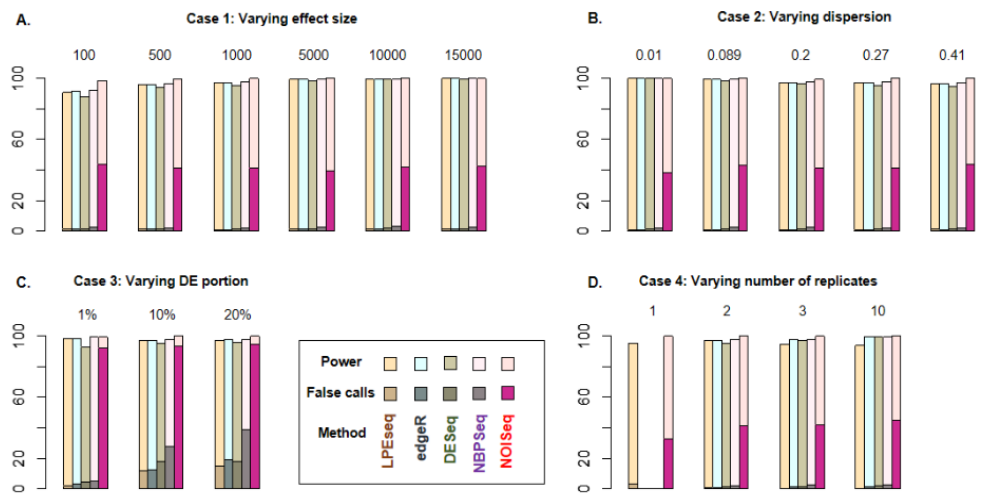


Figure 3-10 Method comparison using simulated data. Power and false discovery rate of each method are denoted as a percentage value in y-axis. The brighter (darker) coloured bars represents the power (false discovery rate) in each case. The denoted values under each subtitle in are read count differences between classes (A), dispersion parameters in negative binomial distribution (B), DE percentages of the total transcripts (C) and numbers of replicates in each class (D).

Firstly, detecting power of all methods increased as the effect size increases (Figure 3-10A). For small effect size, DESeq showed relatively lower power compared to other methods. But it became comparable as effect size increases, while not increasing false discovery. For the effect sizes above 5000, the power of LPEseq almost approached to 100% with the lowest false discovery rate among all methods (Figure 3-10A). Figure 3-10B shows performances of the methods with varying dispersion. As shown in the figures, the power was slightly reduced in all methods (Figure 3-10B). In general, however, all methods showed consistently higher

performances (over 95% of power and below 5% of false calls) regardless of dispersion (Figure 3-10B). The steepest drop in power, made by DESeq was less than 5% among comparisons. Next, DE portion did not much affect the power in all methods, except DESeq (Figure 3-10C). For small number of true DE transcripts, DESeq lost its power, while others remained approximately the same. However, all methods sharply increased in their false discovery rates when DE portion gets large and the minimum increment was achieved by LPEseq (Figure 3-10C).

To further investigate the performance of the methods under different number of replicates available, we varied the number of replicates per each condition. We observed that, in general, when replicates are not available or only a small number of replicates are available, LPEseq provides better performance among the methods (Figure 3-10D). It was similarly observed with MAQC dataset analyses that DESeq lost it power dramatically when applied to non-replicated dataset, while LPEseq were consistent.

For the study of non-replicated data, DESeq only showed comparable performance in limited 11 situations that a small number of DE (1%) has small effect sizes (100, 500 and 1000) with less variety (0.01 and 0.089 dispersion) among 90 simulation situation (data not shown). LPEseq and NOISeq showed over 90% of power in most of our simulated situations with non-replicated data analyses, In case of false discovery, there was not a steady winner. In some situations, LPEseq better performed than NOISeq and vice versa. In particular, LPEseq showed better performed in the situations of smaller effect sizes with large dispersion, while

NOISeq performed well with small dispersion. However, there were some situations where both LPEseq and NOISeq failed to conserve false discovery rate, in particular with large effect size. When a number of replicates are large enough, however, a 3 % of reduction (97% to 94%) in power was observed with LPEseq, while other methods increased to almost 100% in power.

3.6 Conclusion & Discussion

We proposed a method for estimating gene specific variances, especially when only non-replicated datasets are available. By extending LPE method, the proposed method is applicable to the RNA-seq experiments either with or without replicates per each class. Even though the method followed the idea of LPE method, that is, pooling the genes with similar intensity, it is different from LPE in two aspects; (i) estimating local pooled variance from different classes regarding them as replicates (called it 'pseudo-replicates' assumption) and (ii) removing outliers derived from the pseudo-replicates assumption. These two differences make DE analyses with non-replicated datasets feasible. By adding an auxiliary step, our proposed method estimates variance more robustly and reliably even when non-replicated dataset is available.

However, it is worth noting that even though a statistical testing is possible with data without replicates, but the scope of any conclusions drawn from it may be limited and need to be interpreted with extra cautions. But it is also true that we are likely faced with such data, which have a very small number of replicates. In such cases, this study shows that the proposed method is useful in order to obtain the results that are comparable to those obtained from the data with replicates.

In this study, we only considered testing between two varieties in RNA-Seq experiments. However, the idea can be easily extended. For more complicated models, such as analysis of variance (ANOVA), the tests under the 'local pooled

error' and 'pseudo-replicates' assumptions are identical to those in the classical ANOVA case, with the exceptions that the residual mean sum of squares and residual degrees of freedom are evaluated using observations without outliers in a local bin. For other high-throughput experiments, measuring protein expression or DNA methylation, the method can by also applicable to test differential expression.

CHAPTER 4

CONCLUDING REMARKS

Two advances are made in analyzing small sample RNA-Seq data: a new strategy for prioritizing causative variants in a Mendelian family with whole exome sequencing data and a new method for a differential expression analysis with non-replicated RNA sequencing data. More cautions should be taken from designing an experiments and analyzing data from the designs with a limited cost, which cannot afford a large sample.

In such cases, I suggested that Mendelian family is a good research subject and analyzing whole exome sequencing in multiphasic manner that I proposed reduces validation candidates of potential causative variants efficiently. Of course, SNP-chip has been a cost-effective and accurate golden standard technology and used to validate the variants called from sequencing technology in earlier era of NGS. However, it cannot provide personal variants like SNV or CNV, other form of variants. Moreover, sequencing technologies and variant calling steps has become exquisite and stabilized. Thus these two technologies has held firm to their positions in genome studies.

RNA-Seq is another heritage of sequencing technology. Similar in exome sequencing and SNP-chip comparison, RNA-Seq provides more information of the samples under investigation at the price of relatively higher cost. Thus it will more likely lead studies with small samples. I developed a method, called ‘LPEseq’, designed for this purpose. This method is implemented in R language and easy-to-use.

One might think that small sample analysis will be unnecessary when sequencing cost dropped. But I have a feeling of doubt and uncertainty on that opinion. No matter how much the cost will drop, there still needs for small sample analysis because of the limited specimen availability and the pursuit of efficiency and individual-specific information.

APPENDIX A

A-1 R package LPEseq

```
# Package: LPEseq (Differential Expression TEst of RNA-Sequencing data)
# Wrote by Gim, Jungsoo (iedenkim@gmail.com)
cat(paste("#####################################################################
####", "\n"))
cat(paste("\t", "\t", "You are loading LPEseq v1.03", "\n", "\n", sep=""))
cat(paste("\t", "\t", "\t", "Please enjoy...", "\n", "\n", sep=""))
cat(paste("Please visit http://bibs.snu.ac.kr/software/LPEseq for further
info", "\n", sep=""))
cat(paste("#####################################################################
####", "\n"))

## Fine tunning

LPEseq.test <- function(expr_x, expr_y, n.bin=100, df=3, outlier.type="mad",
outlier.pvalue=0.1, k=0.5) {
  n.bin = n.bin
  df = df
  fudge.factor = k
  outlier.type = outlier.type
  outlier.pvalue = outlier.pvalue
  n.x = ncol(as.data.frame(expr_x))
  n.y = ncol(as.data.frame(expr_y))
  n.gene = nrow(as.data.frame(expr_x))
  if(n.x == 1 & n.y == 1){
    mu.x <- expr_x
    mu.y <- expr_y
    tmp_dat <- cbind(expr_x, expr_y)
    tmp.var <- LPEseq.var(tmp_dat, n.bin=n.bin, outlier.type=outlier.type,
pval=outlier.pvalue)
```

```
   var.x <- fudge.factor*LPEseq.predict.var(mu.x, tmp.var, df=df, tol=1e-
6*IQR(tmp.var$mean))
   var.y <- fudge.factor*LPEseq.predict.var(mu.y, tmp.var, df=df, tol=1e-
6*IQR(tmp.var$mean))
   std.dev <- sqrt(var.x + var.y)
   z.stats <- (mu.y-mu.x)/std.dev
   p.val <- as.numeric(2*(1-pnorm(abs(z.stats))))
   adj.p.fdr <- p.adjust(p.val, method="fdr")
   data.out <- data.frame(
     mu.x = mu.x,
     mu.y = mu.y,
     pooled.std.dev = std.dev,
     z.stats = z.stats,
     p.value = p.val,
     q.value = adj.p.fdr
   )
   return(data.out)
  }
  if(n.x == 1 & n.y == 2){
   mu.x <- expr_x
   mu.y <- apply(expr_y, 1, mean)
   tmp.dat <- cbind(mu.x, mu.y)
   tmp.var <- LPEseq.var(tmp_dat, n.bin=n.bin, outlier.type=outlier.type,
pval=outlier.pvalue)
   var.x <- fudge.factor*LPEseq.predict.var(mu.x, tmp.var, df=df, tol=1e-
6*IQR(tmp.var$mean))
   basevar.y <- lpe.var(expr_y, n.bin=n.bin)
   var.y <- basevar.y[,2]
   std.dev <- sqrt(var.x/2 + var.y/2)
   z.stats <- (mu.y-mu.x)/std.dev
   p.val <- as.numeric(2*(1-pnorm(abs(z.stats))))
   adj.p.fdr <- p.adjust(p.val, method="fdr")
     data.out <- data.frame(
     mu.x = mu.x,
     mu.y = mu.y,
     pooled.std.dev = std.dev,
     z.stats = z.stats,
     p.value = p.val,
     q.value = adj.p.fdr
   )
```

```
    return(data.out)
  }
  if(n.x == 1 & n.y > 2){
    effi.y <- 4*n.y/(pi*(2*n.y+1))
    mu.x <- expr_x
    mu.y <- apply(expr_y, 1, mean)
    tmp.dat <- cbind(mu.x, mu.y)
    tmp.var <- LPEseq.var(tmp_dat, n.bin=n.bin, outlier.type=outlier.type,
pval=outlier.pvalue)
    var.x <- fudge.factor*LPEseq.predict.var(mu.x, tmp.var, df=df, tol=1e-
6*IQR(tmp.var$mean))
    basevar.y <- lpe.var(expr_y, n.bin=n.bin)
    var.y <- basevar.y[,2]
    std.dev <- sqrt(var.x/2 + (pi/2)^2*var.y/2)
    z.stats <- (mu.y-mu.x)/std.dev
    p.val <- as.numeric(2*(1-pnorm(abs(z.stats))))
    adj.p.fdr <- p.adjust(p.val, method="fdr")
    data.out <- data.frame(
      mu.x = mu.x,
      mu.y = mu.y,
      pooled.std.dev = std.dev,
      z.stats = z.stats,
      p.value = p.val,
      q.value = adj.p.fdr
    )
    return(data.out)
  }
  if(n.x == 2 & n.y == 1){
    basevar.x <- lpe.var(expr_x, n.bin=n.bin)
    var.x <- basevar.x[,2]
    mu.x <- basevar.x[,1]
    mu.y <- expr_y
    tmp.dat <- cbind(mu.x, mu.y)
    tmp.var <- LPEseq.var(tmp_dat, n.bin=n.bin, outlier.type=outlier.type,
pval=outlier.pvalue)
    var.y <- fudge.factor*LPEseq.predict.var(mu.y, tmp.var, df=df, tol=1e-
6*IQR(tmp.var$mean))
    std.dev <- sqrt(var.x/2 + var.y/2)
    z.stats <- (mu.y-mu.x)/std.dev
    p.val <- as.numeric(2*(1-pnorm(abs(z.stats))))
```

```
  adj.p.fdr <- p.adjust(p.val, method="fdr")
  data.out <- data.frame(
    mu.x = mu.x,
    mu.y = mu.y,
    pooled.std.dev = std.dev,
    z.stats = z.stats,
    p.value = p.val,
    q.value = adj.p.fdr
  )
  return(data.out)
}
if(n.x == 2 & n.y == 2){
  basevar.x <- lpe.var(expr_x, n.bin=n.bin)
  basevar.y <- lpe.var(expr_y, n.bin=n.bin)
  var.x <- basevar.x[,2]
  var.y <- basevar.y[,2]
  mu.x <- basevar.x[,1]
  mu.y <- basevar.y[,1]
  std.dev <- sqrt((var.x/2)+(var.y/2))
  z.stats <- (mu.y-mu.x)/std.dev
  p.val <- as.numeric(2*(1-pnorm(abs(z.stats))))
  adj.p.fdr <- p.adjust(p.val, method="fdr")
  data.out <- data.frame(
    mu.x = mu.x,
    mu.y = mu.y,
    pooled.std.dev = std.dev,
    z.stats = z.stats,
    p.value = p.val,
    q.value = adj.p.fdr
  )
  return(data.out)
}
if(n.x == 2 & n.y > 2){
  basevar.x <- lpe.var(expr_x, n.bin=n.bin)
  var.x <- basevar.x[,2]
  mu.x <- basevar.x[,1]
  basevar.y <- lpe.var(expr_y, n.bin=n.bin)
  mu.y <- basevar.y[,1]
  var.y <- basevar.y[,2]
  std.dev <- sqrt(var.x/2 + (pi/2)^2*var.y/2)
```

```
    z.stats <- (mu.y-mu.x)/std.dev
    p.val <- as.numeric(2*(1-pnorm(abs(z.stats))))
    adj.p.fdr <- p.adjust(p.val, method="fdr")
    data.out <- data.frame(
      mu.x = mu.x,
      mu.y = mu.y,
      pooled.std.dev = std.dev,
      z.stats = z.stats,
      p.value = p.val,
      q.value = adj.p.fdr
    )
    return(data.out)
  }
  if(n.x > 2 & n.y == 1){
    basevar.x <- lpe.var(expr_x, n.bin=n.bin)
    var.x <- basevar.x[,2]
    mu.x <- basevar.x[,1]
    mu.y <- expr_y
    tmp.dat <- cbind(mu.x, mu.y)
    tmp.var  <-  LPEseq.var(tmp_dat,  n.bin=n.bin,  outlier.type=outlier.type,
pval=outlier.pvalue)
    var.y  <-  fudge.factor*LPEseq.predict.var(mu.y,  tmp.var,  df=df,  tol=1e-
6*IQR(tmp.var$mean))
    std.dev <- sqrt((pi/2)^2*var.x/2 + var.y/2)
    z.stats <- (mu.y-mu.x)/std.dev
    p.val <- as.numeric(2*(1-pnorm(abs(z.stats))))
    adj.p.fdr <- p.adjust(p.val, method="fdr")
    data.out <- data.frame(
      mu.x = mu.x,
      mu.y = mu.y,
      pooled.std.dev = std.dev,
      z.stats = z.stats,
      p.value = p.val,
      q.value = adj.p.fdr
    )
    return(data.out)
  }
  if(n.x > 2 & n.y == 2){
    basevar.x <- lpe.var(expr_x, n.bin=n.bin)
    var.x <- basevar.x[,2]
```

```
    mu.x <- basevar.x[,1]
    basevar.y <- lpe.var(expr_y, n.bin=n.bin)
    mu.y <- basevar.y[,1]
    var.y <- basevar.y[,2]
    std.dev <- sqrt((pi/2)^2*var.x/2 + var.y/2)
    z.stats <- (mu.y-mu.x)/std.dev
    p.val <- as.numeric(2*(1-pnorm(abs(z.stats))))
    adj.p.fdr <- p.adjust(p.val, method="fdr")
    data.out <- data.frame(
      mu.x = mu.x,
      mu.y = mu.y,
      pooled.std.dev = std.dev,
      z.stats = z.stats,
      p.value = p.val,
      q.value = adj.p.fdr
    )
    return(data.out)
  }
  if(n.x > 2 & n.y > 2){
    basevar.x <- lpe.var(expr_x, n.bin=n.bin)
    basevar.y <- lpe.var(expr_y, n.bin=n.bin)
    var.x <- basevar.x[,2]
    var.y <- basevar.y[,2]
    mu.x <- basevar.x[,1]
    mu.y <- basevar.y[,1]
    std.dev <- sqrt((pi/2)^2*var.x/2 + (pi/2)^2*var.y/2)
    z.stats <- (mu.y-mu.x)/std.dev
    p.val <- as.numeric(2*(1-pnorm(abs(z.stats))))
    adj.p.fdr <- p.adjust(p.val, method="fdr")
    data.out <- data.frame(
      mu.x = mu.x,
      mu.y = mu.y,
      pooled.std.dev = std.dev,
      z.stats = z.stats,
      p.value = p.val,
      q.value = adj.p.fdr
    )
    return(data.out)
  }
}
```

```
LPEseq.normalise <- function(expr_dat, method="mean"){
  colSum <- apply(expr_dat, 2, sum)
  if(method=="mean"){
    meanVec <- colSum/mean(colSum)
  }else if(method=="median"){
    meanVec <- colSum/median(colSum)
  }else{
     stop("method should be one of \"mean\" or \"median\"")
  }
  normData <- expr_dat
  for(i in 1:ncol(expr_dat)){
    normData[,i] <- expr_dat[,i]/meanVec[i]
  }
  return(log2(normData+1))
}


LPEseq.matrans <- function(dat){
  if(ncol(dat)!=2){
    stop("The number of column should be two")
  }
  MA <- dat
  colnames(MA) = c("A", "M")
  MA[,1] <- apply(dat, 1, mean)
  MA[,2] <- dat[,1] - dat[,2]
  return(MA)
}


am.trans <- function (y){
  if (ncol(y) > 5)
    y <- y[, sample(1:ncol(y), 5)]
  n <- ncol(y)
  if (n < 2) {
    stop("There are no replicated arrays!")
  }
  A <- c()
  M <- c()
```

```
  cc <- permute(1:n)
  for (i in 1:(n - 1)) {
    A <- c(A, c((y + y[, cc[i, ]])/2), recursive = TRUE)
    M <- c(M, c(y - y[, cc[i, ]]), recursive = TRUE)
  }
  return(cbind(A, M))
}


permute <- function(a){
  aa <- matrix(NA, length(a) - 1, length(a))
  for (i in 1:(length(a) - 1)) {
    aa[i, ] <- a[c((i + 1):length(a), 1:i)]
  }
  return(aa)
}


LPEseq.outlier <- function(x, type=c("z", "t", "mad")) {
        n <- length(x)
        s <- match.arg(type)
        ty <- switch(s, z=0, t=1, mad=2)
        if(ty==0){
                res <- (x-mean(x))/sd(x)
                pnorm(res)
        }
        else if(ty==1){
                t <- (x-mean(x))/sd(x)
                res <- (t*sqrt(n-2))/sqrt(n-1-t^2)
                pt(res, n-2)
        }
        else if(ty==2){
                res <- (x - median(x))/mad(x)
                pnorm(res)
        }
}


LPEseq.var <- function(expr_dat, n.bin=100, outlier.type=type, pval=0.05){
  tmp.ma <- LPEseq.matrans(expr_dat)
```

```
  #tmp.inc <- diff(range(tmp.ma[,1]))/n.bin
  #tmp.bin.vec <- seq(range(tmp.ma[,1])[1], range(tmp.ma[,1])[2], tmp.inc)
  tmp.inc <- diff(range(expr_dat[,1]))/n.bin
  tmp.bin.vec <- seq(range(expr_dat[,1])[1], range(expr_dat[,1])[2], tmp.inc)
  tmp.bin.mat <- as.data.frame(matrix(0, nrow=n.bin, ncol=4))
  colnames(tmp.bin.mat) <- c("mean", "var", "ix.NA", "#ofGenes")
  for(i in 1:n.bin){
    tmp.var <- tmp.ma[tmp.ma[,1] >= tmp.bin.vec[i] & tmp.ma[,1] <=
tmp.bin.vec[i+1], 2]
    tmp.bin.mat[i,1] <- mean(tmp.bin.vec[i], tmp.bin.vec[i+1])
    tmp.bin.mat[i,2]       <-       var(tmp.var[which(LPEseq.outlier(tmp.var,
type=outlier.type) >= pval/2 & LPEseq.outlier(tmp.var, type=outlier.type) <=
(1-pval/2))])
    tmp.bin.mat[i,4] <- length(tmp.var)
  }
  tmp.ix.NA <- which(is.na(tmp.bin.mat$var)==F)
  tmp.bin.mat[tmp.ix.NA,3] <- 1
  tmp.bin.mat$var[is.na(tmp.bin.mat$var)] <- 0
  return(tmp.bin.mat)
}


LPEseq.predict.var <- function(gene_expr, qnt_var_mat, df=4, tol = 1e-6 *
IQR(qnt_var_mat$mean)){
  tmp.predict <- numeric(length(gene_expr))
  tmp.pdf      <-      smooth.spline(qnt_var_mat$mean,      qnt_var_mat$var,
w=qnt_var_mat$ix.NA, df=df, tol=tol)
    for(i in 1:length(gene_expr)){
    tmp.predict[i] <- predict(tmp.pdf, gene_expr[i])$y
    if(any(tmp.predict[i]<0)){
      tmp.predict[i] <- min(qnt_var_mat$var.x)
    }
  }
  return(tmp.predict)
}


lpe.var <- function(y, n.bin=n.bin, s.df=10){
  qnt <- 1/n.bin
  AM <- am.trans(y)
```

```
  A <- AM[,1]
  M <- AM[,2]
  median.y <- apply(y, 1, median)
  quantile.A <- quantile(A, probs = seq(0,1,qnt), na.rm=T)
  quan.n <- length(quantile.A)-1
  var.M <- rep(0, length=quan.n-1)
  medianAs <- rep(0, length=quan.n-1)
  if(sum(A==min(A)) > (qnt*length(A))){
    tmpA <- A[!(A==min(A))]
    quantile.A <- c(min(A), quantile(tmpA, probs=seq(qnt, 1, qnt), na.rm=T))
  }
  for(i in 2:(quan.n+1)){
    n.i <- length(!is.na(M[A>=quantile.A[i-1]&A<quantile.A[i]]))
    if(n.i > 1){
      mult.factor <- 0.5*((n.i-0.5)/(n.i-1))
      var.M[i-1] <- mult.factor * var(M[A>=quantile.A[i-1] & A<quantile.A[i]],
na.rm=T)
      medianAs[i-1] <- median(A[A>=quantile.A[i-1] & A<quantile.A[i]], na.rm=T)
    }
  }
  if(any(is.na(var.M))){
    for(i in (quan.n-1):1){
      if(is.na(var.M[i])){
        var.M[i] <- ifelse(!is.na(var.M[i-1]), mean(var.M[i+1], var.M[i-1]),
var.M[i+1])
      }
    }
  }
  var.M[1:which(var.M==max(var.M))] <- max(var.M)
  base.var <- cbind(A=medianAs, var.M=var.M)
  sm.spline <- smooth.spline(base.var[,1], base.var[,2], df=s.df)
  min.Var <- min(base.var[,2])
  var.genes <- fixbounds.predict.smooth.spline(sm.spline, median.y)$y
  if(any(var.genes < min.Var))
    var.genes[var.genes < min.Var] <- min.Var
  basevar.step1 <- cbind(A=median.y, var.M=var.genes)
#  ord.median <- order(basevar.step1[,1])
#  var.genes.ord <- basevar.step1[ord.median,]
#  return(var.genes.ord)
  return(basevar.step1)
```

```
}



fixbounds.predict.smooth.spline <- function (object, x, deriv = 0){
  if (missing(x)) {
    if (deriv == 0) {
      return(object[c("x", "y")])
    }
    else {
      x <- object$x
    }
  }
  if (is.null(object)) {
    stop("not a valid smooth.spline object")
  }
  else {
    out <- predict(object, x, deriv)
    maxpredY <- object$y[object$x == max(object$x)]
    out$y[out$x > max(object$x)] <- maxpredY
    minpredY <- object$y[object$x == min(object$x)]
    out$y[out$x < min(object$x)] <- minpredY
    invisible(out)
  }
}



generateData <- function(n.gene=20000, n.cond=2, n.deg=0, eff=1000, n.rep=3,
disp=0.25){
  tmp.size <- 1/disp
  tmp.ix.small <- which(para.maxLik[,1] < 10)
  tmp.ix.large <- which(para.maxLik[,1] >= 10)
  tmp.mu1 <- para.maxLik[sample(tmp.ix.small, n.gene/2, replace=T),1]
  tmp.mu2 <- para.maxLik[sample(tmp.ix.large, n.gene/2, replace=T),1]
  tmp.mu <- c(tmp.mu1, tmp.mu2)
  tmp.dat <- matrix(0, nrow=n.gene, ncol=n.cond*n.rep+1)
  for(i in 1:n.gene){
    tmp.dat[i,1:(n.cond*n.rep)]   <-   rnbinom(n.cond*n.rep,   size=tmp.size,
mu=tmp.mu[i])
  }
```

```
  if(n.deg!=0){
    deg.ix <- sample(1:n.gene, n.deg)
    deg.ix.u <- sample(deg.ix, round(n.deg/2))
    deg.ix.l <- deg.ix[!deg.ix %in% deg.ix.u]
    for(i in 1:length(deg.ix.u)){
      tmp.dat[deg.ix.u[i],    (n.rep+1):(n.cond*n.rep)]    <-    rnbinom(n.rep,
size=tmp.size, mu=tmp.mu[deg.ix.u[i]]+eff)
    }
    for(i in 1:length(deg.ix.l)){
      tmp.dat[deg.ix.l[i],    1:n.rep]    <-    rnbinom(n.rep,    size=tmp.size,
mu=tmp.mu[deg.ix.l[i]]+eff)
    }
    tmp.dat[deg.ix, n.cond*n.rep+1] <- 1
  }
  rowVec <- paste("gene", 1:n.gene, sep="_")
  colVec  <-  c(paste("condition1",  1:n.rep,  sep="."),  paste("condition2",
1:n.rep, sep="."), "DEG")
  rownames(tmp.dat) <- rowVec
  colnames(tmp.dat) <- colVec
  return(tmp.dat)
}
```

A-2 LPEseq manual

`generateData(n.gene, n.cond, n.deg, eff, n.rep, disp)`

Generating simulation data

Arguments

n.gene:	the number of genes (default 20 000)
n.cond:	the number of experimental conditions (default 2)
n.deg:	the number of differentially expressed genes (default 0)
eff:	count difference between differential expression (default 1000)
n.rep:	the number of replicates per each condition (default 3)
disp:	dispersion parameter, inverse of 'size' parameter for nbinom-function (default 0.25; biological replicates, 0.01)

`LPEseq.normalise(data, method)`

Normalising input data according to their total count values

Arguments

data: count value matrix

method: c("mean", "median") for summary of the column sums, default: “mean”

`LPEseq.matrans(dat)`

MA transformation of the data

Arguments

dat: expression data with 2 columns

`LPEseq.outlier(x, type=c("z", "t", "mad")`

Returns p-values of outlying observations

Arguments

x:	numeric vector
type:	one of c("z", "t", "mad"), "mad" is recommended

`LPEseq.var(expr_dat, n.bin, outlier.type, pval)`

Evaluates local pooled variance

Arguments

expr_dat: normalized expression data with 2 columns
n.bin: the number of bins (default = 100)
outlier.type: argument used in LPEseq.outlier. One of ("z", "t", "mad").
pval: pvalue threshold for outlier filtering (default: 0.1)

`LPEseq.predict.var(gene_expr, qnt_var_mat, df, tol)`

Function predicting per gene variance using local pooled variance curve
Arguments
gene_expr: gene expression whose variance to be estimated
qnt_var_mat: output of LPEseq.var function
df: smoothing degree (default: 3)
tol: tolerance parameter in smooth.spline function

`LPEseq.test(expr_x, expr_y, n.bin, df, outlier.type, pval, k)`

Differential expression test
Arguments
expr_x: a numeric value of the first condition
expr_y: a numeric value of the second condition
n.bin: output of plpv.permutedVariance function
df: the desired equivalent number of d.o.f (trace of the smoother matrix; default 3)
outlier.type: c("z", "t", "mad") for outlier detection
pval: threshold for removing outlier (default: 0.1)
k: fudge factor (efficacy value, pi/2 for using median in original LPE method)

A-3 Example script

The package LPEseq provides a method to test for differential expression analysis for RNA-Seq data with no replicate. Although, many of the methods are available including EdgeR, DESeq, CuffDiff, etc., only few methods deal with non-replicated data properly. This brief script is written for the users to explain how to use the LPEseq package. Anyone who are interested in detailed method, please download and see our manuscript on http://bibs.snu.ac.kr/software/LPEseq.

1. Input Data: generation

LPEseq starts its analysis with read counts. Therefore you have to equip yourself with RNA-Seq read count data sets on your hand first. If you are not familiar with that, please visit web-sites to learn how to obtain such a data. Good references are *GenomicRanges* in Bioconductor, *htseq-count* script written in Python framework, *Cufflinks*, the well-known software, etc. For this LPEseq usage example, we do not need count dataset. Using `generateData()`, you can learn how to use LPEseq package without any real count dataset.

The most recent version of LPEseq package 'LPEseq_version.R' should be downloaded before going further. If you downloaded already, just run the source code.

```
> source("LPEseq_version.R")
############################################################
```

```
                You are loading LPEseq v1.03


                      Please enjoy...


Please visit http://bibs.snu.ac.kr/software/LPEseq for
further info
#################################################################
```

Or you can download it directly

```
>
source("http://bibs.snu.ac.kr/software/LPEseq/LPEseq_v1_03.R"
)
#################################################################
                You are loading LPEseq v1.03


                      Please enjoy...


Please visit http://bibs.snu.ac.kr/software/LPEseq for further
info
#################################################################
```

Now you are ready to generate the simulation datasets by typing

```
> simData <- generateData()
```

In case of you have your own count data set, you can read the data with read.table()

```
> yourData <- read.table("your_data.txt", header = , sep = ,
…)
```

Once you loaded your own data, further analysis procedure is the same. Since the data is generated to follow negative binomial distribution, you can observe this pattern

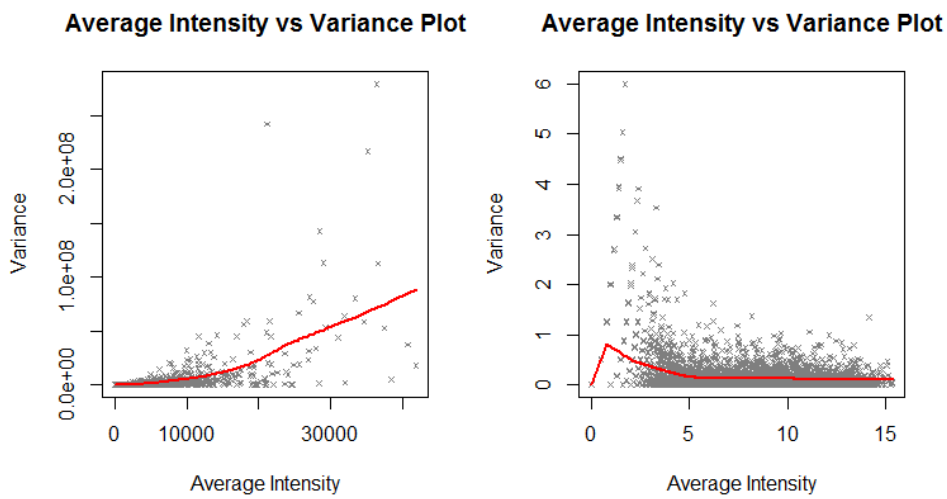


Figure A-1 Variance versus mean intensity plot with original intensity (left) and log-transformed intensity (right)

2. Normalization

To get rid of so-called "size factors", LPEseq follows the similar method of DESeq's, which divide each column of the count table by the size factor for this column. By doing so, the count values are brought to a common comparable scale. The difference is that LPEseq adds pseudo-count value 1 to all the values and take log-2 transformation. Before going further, let's take `head()` to the generated simulation

data.

```
> head(simData)
```

```
      condition1.1 condition2.1 DEG
gene_1           0           0   0
gene_2           0           0   0
gene_3           0           0   0
gene_4           0           1   0
gene_5           0           0   0
gene_6           0           0   0
```

As we might see, `generatedData()` function generate DEG index at the end of the column. Therefore, when normalizing simulation data, include only count values in an argument,

```
> simData.norm <- LPEseq.normalise(simData[,-3])
```

If your own data consists of original count values, exactly the same script will do,

```
> youData.norm <-
LPEseq.normalise(yourData_if_original_count)
```

But if your data includes normalized count values, such RPKM or FPKM, just take log-transformation to your data, then that will conduct the same normalization procedure.

```
> youData.norm <- log(yourData_if_FPKM_value, base = 2)
```

Or I recommend performing `LPEseq.normalise()` again on your normalized

data.

```
> head(simData.norm)

condition1.1 condition2.1
gene_1          0    0.0000000
gene_2          0    0.0000000
gene_3          0    0.0000000
gene_4          0    0.9976644
gene_5          0    0.0000000
gene_6          0    0.0000000
```

3. Evaluating LPE variance

Now you are at the stage of calculating LPE estimator. There're several arguments: *n.bin, outlier.type* and *pval*. Argument *n.bin* is simply the desired number of quantiling bin with default value of 100 (recommended). Argument *outlier.type* is one of either "*z"* , *"t"* and *"mad"*. Default is "*mad*", and I recommend for the users to use *"mad"* for most of the data. The last argument *pval* is the p-value threshold for indicating outliers and removing them in calculation. 0.1 is recommended for the *pval.*

```
> simData.norm.var <- LPEseq.var(simData.norm, n.bin=100,
outlier.type="mad", pval=0.1)

> head(simData.norm.var)

      mean var ix.NA #ofGenes
1 0.0000000   0     0     9162
2 0.1598984   0     0        0
```

```
3 0.3197969   0     0       0
4 0.4796953   0     0     242
5 0.6395937   0     0      57
6 0.7994921   0     0       0
```

`LPEseq.var()` returns mean intensity and variance and index for NA and the number of genes in each quantile bin.

4. DEG test

The procedure for DEG analysis in LPEseq is straightforward. `LPEseq.test()` function can be directly applied in normalized count matrix

```
> simData.res <- LPEseq.test(simData.norm[,1],
simData.norm[,2])

> head(simData.res)

       mu.x     mu.y pooled.std.dev  z.stats   p.value q.value
gene_1    0 0.0000000      1.334506 0.000000 1.0000000       1
gene_2    0 0.0000000      1.334506 0.000000 1.0000000       1
gene_3    0 0.0000000      1.334506 0.000000 1.0000000       1
gene_4    0 0.9976644      1.292988 0.771596 0.4403538       1
gene_5    0 0.0000000      1.334506 0.000000 1.0000000       1
gene_6    0 0.0000000      1.334506 0.000000 1.0000000       1
```

Since `LPEseq.test()` reports simply the nominal p-value and q.value (adjusted by `p.adjust(, method="BH")`), we can find and report information about up-regulated and down-regulated genes using following function,

```
> deg_list <- function(data, pval.vec, pval.threshold){
       deg.ix <- which(pval.vec < pval.threshold)
```

```
        up.deg.ix <- deg.ix[apply(data[deg.ix,], 1, diff) >=0]
        lo.deg.ix <- deg.ix[apply(data[deg.ix,], 1, diff) < 0]
        upgenes <- cbind(rownames(data[up.deg.ix,]),
pval.vec[up.deg.ix])
        colnames(upgenes) <- c("GeneName", "p_value")
        rownames(upgenes) <- paste(1:nrow(upgenes))
        logenes <- cbind(rownames(data[lo.deg.ix,]),
pval.vec[lo.deg.ix])
        colnames(logenes) <- c("GeneName", "p_value")
        rownames(logenes) <- paste(1:nrow(logenes))
        return(list(up.genes = upgenes, lo.genes = logenes))
    }
> tmpRes <- deg_list(simData, res.ma[,1], 0.05)
> names(tmpRes)
[1] "u.gene"  "l.genes"
> head(tmpRes$up.genes)
GeneName   p_value
1 "gene 68"  "0.023693536255571"
2 "gene 111" "0.0435464965693764"
3 "gene 122" "0.023693536255571"
4 "gene 144" "0.023693536255571"
5 "gene 203" "0.00332397439981746"
6 "gene 224" "0.0439493021943069"
```

5. Session Info

```
> sessionInfo()

R version 3.0.1 (2013-05-16)
Platform: i386-w64-mingw32/i386 (32-bit)

locale:
[1] LC_COLLATE=Korean_Korea.949  LC_CTYPE=Korean_Korea.949
[3] LC_MONETARY=Korean_Korea.949 LC_NUMERIC=C
[5] LC_TIME=Korean_Korea.949

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods
base
```

Bibliography

Abou Jamra, R., O. Philippe, A. Raas-Rothschild, S. H. Eck, E. Graf, R. Buchert, G. Borck, A. Ekici, F. F. Brockschmidt, M. M. Nothen, A. Munnich, T. M. Strom, A. Reis and L. Colleaux (2011). "Adaptor protein complex 4 deficiency causes severe autosomal-recessive intellectual disability, progressive spastic paraplegia, shy character, and short stature." Am J Hum Genet **88**(6): 788-795.

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov and S. R. Sunyaev (2010). "A method and server for predicting damaging missense mutations." Nat Methods **7**(4): 248-249.

Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data." Genome Biol **11**(10): R106.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Baek, J. I., S. K. Oh, D. B. Kim, S. Y. Choi, U. K. Kim, K. Y. Lee and S. H. Lee (2012). "Targeted massive parallel sequencing: the effective detection of novel causative mutations associated with hearing loss in small families." Orphanet J Rare Dis **7**: 60.

Bamshad, M. J., S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson and J. Shendure (2011). "Exome sequencing as a tool for Mendelian disease gene discovery." Nat Rev Genet **12**(11): 745-755.

Belyantseva, I. A., B. J. Perrin, K. J. Sonnemann, M. Zhu, R. Stepanyan, J. McGee, G. I. Frolenkov, E. J. Walsh, K. H. Friderici, T. B. Friedman and J. M. Ervasti (2009). "Gamma-actin is required for cytoskeletal maintenance but not development." Proc Natl Acad Sci U S A **106**(24): 9703-9708.

Bilguvar, K., A. K. Ozturk, A. Louvi, K. Y. Kwan, M. Choi, B. Tatli, D. Yalnizoglu, B. Tuysuz, A. O. Caglayan, S. Gokben, H. Kaymakcalan, T. Barak, M. Bakircioglu, K. Yasuno, W. Ho, S. Sanders, Y. Zhu, S. Yilmaz, A. Dincer, M. H. Johnson, R. A. Bronen,

N. Kocer, H. Per, S. Mane, M. N. Pamir, C. Yalcinkaya, S. Kumandas, M. Topcu, M. Ozmen, N. Sestan, R. P. Lifton, M. W. State and M. Gunel (2010). "Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations." Nature **467**(7312): 207-210.

Bolze, A., M. Byun, D. McDonald, N. V. Morgan, A. Abhyankar, L. Premkumar, A. Puel, C. M. Bacon, F. Rieux-Laucat, K. Pang, A. Britland, L. Abel, A. Cant, E. R. Maher, S. J. Riedl, S. Hambleton and J. L. Casanova (2010). "Whole-exome-sequencing-based discovery of human FADD deficiency." Am J Hum Genet **87**(6): 873-881.

Bowden, D. W., S. S. An, N. D. Palmer, W. M. Brown, J. M. Norris, S. M. Haffner, G. A. Hawkins, X. Guo, J. I. Rotter, Y. D. Chen, L. E. Wagenknecht and C. D. Langefeld (2010). "Molecular basis of a linkage peak: exome sequencing and family-based analysis identify a rare genetic variant in the ADIPOQ gene in the IRAS Family Study." Hum Mol Genet **19**(20): 4112-4120.

Bullard, J. H., E. Purdom, K. D. Hansen and S. Dudoit (2010). "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments." BMC Bioinformatics **11**: 94.

Carelle-Calmels, N., P. Saugier-Veber, F. Girard-Lemaire, G. Rudolf, B. Doray, E. Guerin, P. Kuhn, M. Arrive, C. Gilch, E. Schmitt, S. Fehrenbach, A. Schnebelen, T. Frebourg and E. Flori (2009). "Genetic compensation in a human genomic disorder." N Engl J Med **360**(12): 1211-1216.

Carrasquillo, M. M., A. S. McCallion, E. G. Puffenberger, C. S. Kashuk, N. Nouri and A. Chakravarti (2002). "Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease." Nat Genet **32**(2): 237-244.

Cherny, S. S., G. R. Abecasis, W. O. Cookson, P. C. Sham and L. R. Cardon (2001). "The effect of genotype and pedigree error on linkage analysis: analysis of three asthma genome scans." Genet Epidemiol **21 Suppl 1**: S117-122.

Cho, T. J., O. H. Kim, I. H. Choi, G. Nishimura, A. Superti-Furga, K. S. Kim, Y. J. Lee and W. Y. Park (2010). "A dominant mesomelic dysplasia associated with a 1.0-Mb microduplication of HOXD gene cluster at 2q31.1." J Med Genet **47**(9): 638-639.

Cooper, G. M. and J. Shendure (2011). "Needles in stacks of needles: finding

disease-causal variants in a wealth of genomic data." Nat Rev Genet **12**(9): 628-640.

de Heer, A. M., P. L. Huygen, R. W. Collin, J. Oostrik, H. Kremer and C. W. Cremers (2009). "Audiometric and vestibular features in a second Dutch DFNA20/26 family with a novel mutation in ACTG1." Ann Otol Rhinol Laryngol **118**(5): 382-390.

Di, Y. M., D. W. Schafer, J. S. Cumbie and J. H. Chang (2011). "The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq." Statistical Applications in Genetics and Molecular Biology **10**(1).

Dror, A. A. and K. B. Avraham (2010). "Hearing impairment: a panoply of genes and functions." Neuron **68**(2): 293-308.

Drummond, M. C., I. A. Belyantseva, K. H. Friderici and T. B. Friedman (2012). "Actin in hair cells and hearing loss." Hear Res **288**(1-2): 89-99.

Flanagan, S. E., A. M. Patch and S. Ellard (2010). "Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations." Genet Test Mol Biomarkers **14**(4): 533-537.

Frazee, A. C., B. Langmead and J. T. Leek (2011). "ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets." BMC Bioinformatics **12**: 449.

Gilissen, C., A. Hoischen, H. G. Brunner and J. A. Veltman (2012). "Disease gene identification strategies for exome sequencing." Eur J Hum Genet **20**(5): 490-497.

Gim, J., H. S. Kim, J. Kim, M. Choi, J. R. Kim, Y. J. Chung and K. H. Cho (2010). "A system-level investigation into the cellular toxic response mechanism mediated by AhR signal transduction pathway." Bioinformatics **26**(17): 2169-2175.

Gonzalez-Navarro, F. F. and L. A. Belanche-Munoz (2011). "Parsimonious selection of useful genes in microarray gene expression data." Adv Exp Med Biol **696**: 45-55.

Gregg, C., J. Zhang, B. Weissbourd, S. Luo, G. P. Schroth, D. Haig and C. Dulac (2010). "High-resolution analysis of parent-of-origin allelic expression in the mouse brain." Science **329**(5992): 643-648.

Hammer, P., M. S. Banck, R. Amberg, C. Wang, G. Petznick, S. Luo, I. Khrebtukova, G. P. Schroth, P. Beyerlein and A. S. Beutler (2010). "mRNA-seq with agnostic splice

site discovery for nervous system transcriptomics tested in chronic pain." Genome Res **20**(6): 847-860.
Henningsen, A. and O. Toomet (2011). "maxLik: A package for maximum likelihood estimation in R." Computational Statistics **26**(3): 443-458.
Hilgert, N., R. J. Smith and G. Van Camp (2009). "Forty-six genes causing nonsyndromic hearing impairment: which ones should be analyzed in DNA diagnostics?" Mutat Res **681**(2-3): 189-196.
Hollox, E. J., J. C. Barber, A. J. Brookes and J. A. Armour (2008). "Defensins and the dynamic genome: what we can learn from structural variation at human chromosome band 8p23.1." Genome Res **18**(11): 1686-1697.
Huang, R. S., P. Chen, S. Wisel, S. Duan, W. Zhang, E. H. Cook, S. Das, N. J. Cox and M. E. Dolan (2009). "Population-specific GSTM1 copy number variation." Hum Mol Genet **18**(2): 366-372.
International HapMap, C. (2003). "The International HapMap Project." Nature **426**(6968): 789-796.
International Schizophrenia, C. (2008). "Rare chromosomal deletions and duplications increase risk of schizophrenia." Nature **455**(7210): 237-241.
Jain, E. and K. Jain (2001). "Integrated bioinformatics- high-throughput interpretation of pathways and biology." Trends Biotechnol **19**(5): 157-158.
Jain, N., J. Thatte, T. Braciale, K. Ley, M. O'Connell and J. K. Lee (2003). "Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays." Bioinformatics **19**(15): 1945-1951.
Johnson, J. O., J. Mandrioli, M. Benatar, Y. Abramzon, V. M. Van Deerlin, J. Q. Trojanowski, J. R. Gibbs, M. Brunetti, S. Gronka, J. Wuu, J. Ding, L. McCluskey, M. Martinez-Lage, D. Falcone, D. G. Hernandez, S. Arepalli, S. Chong, J. C. Schymick, J. Rothstein, F. Landi, Y. D. Wang, A. Calvo, G. Mora, M. Sabatelli, M. R. Monsurro, S. Battistini, F. Salvi, R. Spataro, P. Sola, G. Borghero, I. Consortium, G. Galassi, S. W. Scholz, J. P. Taylor, G. Restagno, A. Chio and B. J. Traynor (2010). "Exome sequencing reveals VCP mutations as a cause of familial ALS." Neuron **68**(5): 857-864.
Kalay, E., G. Yigit, Y. Aslan, K. E. Brown, E. Pohl, L. S. Bicknell, H. Kayserili, Y. Li, B.

Tuysuz, G. Nurnberg, W. Kiess, M. Koegl, I. Baessmann, K. Buruk, B. Toraman, S. Kayipmaz, S. Kul, M. Ikbal, D. J. Turner, M. S. Taylor, J. Aerts, C. Scott, K. Milstein, H. Dollfus, D. Wieczorek, H. G. Brunner, M. Hurles, A. P. Jackson, A. Rauch, P. Nurnberg, A. Karaguzel and B. Wollnik (2011). "CEP152 is a genome maintenance protein disrupted in Seckel syndrome." Nat Genet **43**(1): 23-26.

Kang, G. C., A. W. Gan, A. Yam, A. B. Tan and S. C. Tay (2010). "Mycobacterium abscessus Hand Infections in Immunocompetent Fish Handlers: Case Report." J Hand Surg Am **35**(7): 1142-1145.

Khaitlina, S. Y. (2001). "Functional specificity of actin isoforms." Int Rev Cytol **202**: 35-98.

Krawitz, P. M., M. R. Schweiger, C. Rodelsperger, C. Marcelis, U. Kolsch, C. Meisel, F. Stephani, T. Kinoshita, Y. Murakami, S. Bauer, M. Isau, A. Fischer, A. Dahl, M. Kerick, J. Hecht, S. Kohler, M. Jager, J. Grunhagen, B. J. de Condor, S. Doelken, H. G. Brunner, P. Meinecke, E. Passarge, M. D. Thompson, D. E. Cole, D. Horn, T. Roscioli, S. Mundlos and P. N. Robinson (2010). "Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome." Nat Genet **42**(10): 827-829.

Laird, N. M. and C. Lange (2006). "Family-based designs in the age of large-scale gene-association studies." Nat Rev Genet **7**(5): 385-394.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and S. Genome Project Data Processing (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, J., R. Lupat, K. C. Amarasinghe, E. R. Thompson, M. A. Doyle, G. L. Ryland, R. W. Tothill, S. K. Halgamuge, I. G. Campbell and K. L. Gorringe (2012). "CONTRA: copy number analysis for targeted resequencing." Bioinformatics **28**(10): 1307-1313.

Liu, P., H. Li, X. Ren, H. Mao, Q. Zhu, Z. Zhu, R. Yang, W. Yuan, J. Liu, Q. Wang and M. Liu (2008). "Novel ACTG1 mutation causing autosomal dominant non-syndromic hearing impairment in a Chinese family." J Genet Genomics **35**(9): 553-558.

Maher, C. A., C. Kumar-Sinha, X. Cao, S. Kalyana-Sundaram, B. Han, X. Jing, L. Sam, T. Barrette, N. Palanisamy and A. M. Chinnaiyan (2009). "Transcriptome sequencing

to detect gene fusions in cancer." Nature **458**(7234): 97-101.
Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens and Y. Gilad (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." Genome Res **18**(9): 1509-1517.
McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly and M. A. DePristo (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome Res **20**(9): 1297-1303.
Min, B. J., N. Kim, T. Chung, O. H. Kim, G. Nishimura, C. Y. Chung, H. R. Song, H. W. Kim, H. R. Lee, J. Kim, T. H. Kang, M. E. Seo, S. D. Yang, D. H. Kim, S. B. Lee, J. I. Kim, J. S. Seo, J. Y. Choi, D. Kang, D. Kim, W. Y. Park and T. J. Cho (2011). "Whole-exome sequencing identifies mutations of KIF22 in spondyloepimetaphyseal dysplasia with joint laxity, leptodactylic type." Am J Hum Genet **89**(6): 760-766.
Modlich, O., H. B. Prisack, M. Munnes, W. Audretsch and H. Bojar (2005). "Predictors of primary breast cancers responsiveness to preoperative epirubicin/cyclophosphamide-based chemotherapy: translation of microarray data into clinically useful predictive signatures." J Transl Med **3**: 32.
Montgomery, S. B., M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo and E. T. Dermitzakis (2010). "Transcriptome genetics using second generation sequencing in a Caucasian population." Nature **464**(7289): 773-777.
Morell, R. J., K. H. Friderici, S. Wei, J. L. Elfenbein, T. B. Friedman and R. A. Fisher (2000). "A new locus for late-onset, progressive, hereditary hearing loss DFNA20 maps to 17q25." Genomics **63**(1): 1-6.
Morin, M., K. E. Bryan, F. Mayo-Merino, R. Goodyear, A. Mencia, S. Modamio-Hoybjor, I. del Castillo, J. M. Cabalka, G. Richardson, F. Moreno, P. A. Rubenstein and M. A. Moreno-Pelayo (2009). "In vivo and in vitro effects of two novel gamma-actin (ACTG1) mutations that cause DFNA20/26 hearing impairment." Hum Mol Genet **18**(16): 3075-3089.
Musunuru, K., J. P. Pirruccello, R. Do, G. M. Peloso, C. Guiducci, C. Sougnez, K. V. Garimella, S. Fisher, J. Abreu, A. J. Barry, T. Fennell, E. Banks, L. Ambrogio, K.

Cibulskis, A. Kernytsky, E. Gonzalez, N. Rudzicz, J. C. Engert, M. A. DePristo, M. J. Daly, J. C. Cohen, H. H. Hobbs, D. Altshuler, G. Schonfeld, S. B. Gabriel, P. Yue and S. Kathiresan (2010). "Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia." N Engl J Med **363**(23): 2220-2227.

Ng, S. B., A. W. Bigham, K. J. Buckingham, M. C. Hannibal, M. J. McMillin, H. I. Gildersleeve, A. E. Beck, H. K. Tabor, G. M. Cooper, H. C. Mefford, C. Lee, E. H. Turner, J. D. Smith, M. J. Rieder, K. Yoshiura, N. Matsumoto, T. Ohta, N. Niikawa, D. A. Nickerson, M. J. Bamshad and J. Shendure (2010). "Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome." Nat Genet **42**(9): 790-793.

Ng, S. B., K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure and M. J. Bamshad (2010). "Exome sequencing identifies the cause of a mendelian disorder." Nat Genet **42**(1): 30-35.

Ng, S. B., D. A. Nickerson, M. J. Bamshad and J. Shendure (2010). "Massively parallel sequencing and rare disease." Hum Mol Genet **19**(R2): R119-124.

Nielsen, R., J. S. Paul, A. Albrechtsen and Y. S. Song (2011). "Genotype and SNP calling from next-generation sequencing data." Nat Rev Genet **12**(6): 443-451.

Oshlack, A., M. D. Robinson and M. D. Young (2010). "From RNA-seq reads to differential expression results." Genome Biol **11**(12): 220.

Ott, J., Y. Kamatani and M. Lathrop (2011). "Family-based designs for genome-wide association studies." Nat Rev Genet **12**(7): 465-474.

Otto, E. A., T. W. Hurd, R. Airik, M. Chaki, W. Zhou, C. Stoetzel, S. B. Patil, S. Levy, A. K. Ghosh, C. A. Murga-Zamalloa, J. van Reeuwijk, S. J. Letteboer, L. Sang, R. H. Giles, Q. Liu, K. L. Coene, A. Estrada-Cuzcano, R. W. Collin, H. M. McLaughlin, S. Held, J. M. Kasanuki, G. Ramaswami, J. Conte, I. Lopez, J. Washburn, J. Macdonald, J. Hu, Y. Yamashita, E. R. Maher, L. M. Guay-Woodford, H. P. Neumann, N. Obermuller, R. K. Koenekoop, C. Bergmann, X. Bei, R. A. Lewis, N. Katsanis, V. Lopes, D. S. Williams, R. H. Lyons, C. V. Dang, D. A. Brito, M. B. Dias, X. Zhang, J. D. Cavalcoli, G. Nurnberg, P. Nurnberg, E. A. Pierce, P. K. Jackson, C. Antignac, S. Saunier, R. Roepman, H. Dollfus, H. Khanna and F. Hildebrandt (2010). "Candidate exome capture identifies mutation of SDCCAG8 as the cause of a retinal-renal

ciliopathy." Nat Genet **42**(10): 840-850.
Perrin, B. J., K. J. Sonnemann and J. M. Ervasti (2010). "beta-actin and gamma-actin are each dispensable for auditory hair cell development but required for Stereocilia maintenance." PLoS Genet **6**(10): e1001158.
Rendtorff, N. D., M. Zhu, T. Fagerheim, T. L. Antal, M. Jones, T. M. Teslovich, E. M. Gillanders, M. Barmada, E. Teig, J. M. Trent, K. H. Friderici, D. A. Stephan and L. Tranebjaerg (2006). "A novel missense mutation in ACTG1 causes dominant deafness in a Norwegian DFNA20/26 family, but ACTG1 mutations are not frequent among families with hereditary hearing impairment." Eur J Hum Genet **14**(10): 1097-1105.
Risso, D., K. Schwartz, G. Sherlock and S. Dudoit (2011). "GC-content normalization for RNA-Seq data." BMC Bioinformatics **12**: 480.
Riviere, J. B., B. W. van Bon, A. Hoischen, S. S. Kholmanskikh, B. J. O'Roak, C. Gilissen, S. Gijsen, C. T. Sullivan, S. L. Christian, O. A. Abdul-Rahman, J. F. Atkin, N. Chassaing, V. Drouin-Garraud, A. E. Fry, J. P. Fryns, K. W. Gripp, M. Kempers, T. Kleefstra, G. M. Mancini, M. J. Nowaczyk, C. M. van Ravenswaaij-Arts, T. Roscioli, M. Marble, J. A. Rosenfeld, V. M. Siu, B. B. de Vries, J. Shendure, A. Verloes, J. A. Veltman, H. G. Brunner, M. E. Ross, D. T. Pilz and W. B. Dobyns (2012). "De novo mutations in the actin genes ACTB and ACTG1 cause Baraitser-Winter syndrome." Nat Genet **44**(4): 440-444, S441-442.
Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics **26**(1): 139-140.
Robinson, M. D. and A. Oshlack (2010). "A scaling normalization method for differential expression analysis of RNA-seq data." Genome Biol **11**(3): R25.
Robinson, M. D. and G. K. Smyth (2007). "Moderated statistical tests for assessing differences in tag abundance." Bioinformatics **23**(21): 2881-2887.
Rosenthal, E. A., J. Ronald, J. Rothstein, R. Rajagopalan, J. Ranchalis, G. Wolfbauer, J. J. Albers, J. D. Brunzell, A. G. Motulsky, M. J. Rieder, D. A. Nickerson, E. M. Wijsman and G. P. Jarvik (2011). "Linkage and association of phospholipid transfer protein activity to LASS4." J Lipid Res **52**(10): 1837-1846.

Ruano, Y., M. Mollejo, A. R. de Lope, J. L. Hernandez-Moneo, P. Martinez and B. Melendez (2010). "Microarray-based comparative genomic hybridization (array-CGH) as a useful tool for identifying genes involved in Glioblastoma (GB)." Methods Mol Biol **653**: 35-45.

Saito, A. and N. Kamatani (2002). "Strategies for genome-wide association studies: optimization of study designs by the stepwise focusing method." J Hum Genet **47**(7): 360-365.

Schwarz, J. M., C. Rodelsperger, M. Schuelke and D. Seelow (2010). "MutationTaster evaluates disease-causing potential of sequence alterations." Nat Methods **7**(8): 575-576.

Sha, B. Y., T. L. Yang, L. J. Zhao, X. D. Chen, Y. Guo, Y. Chen, F. Pan, Z. X. Zhang, S. S. Dong, X. H. Xu and H. W. Deng (2009). "Genome-wide association study suggested copy number variation may be associated with body mass index in the Chinese population." J Hum Genet **54**(4): 199-202.

Shaffer, C. (2007). "Next-generation sequencing outpaces expectations." Nat Biotechnol **25**(2): 149.

Sirmaci, A., T. Walsh, H. Akay, M. Spiliopoulos, Y. B. Sakalar, A. Hasanefendioglu-Bayrak, D. Duman, A. Farooq, M. C. King and M. Tekin (2010). "MASP1 mutations in patients with facial, umbilical, coccygeal, and auditory findings of Carnevale, Malpuech, OSA, and Michels syndromes." Am J Hum Genet **87**(5): 679-686.

Smith, K. R., C. J. Bromhead, M. S. Hildebrand, A. E. Shearer, P. J. Lockhart, H. Najmabadi, R. J. Leventer, G. McGillivray, D. J. Amor, R. J. Smith and M. Bahlo (2011). "Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes." Genome Biol **12**(9): R85.

Soneson, C. and M. Delorenzi (2013). "A comparison of methods for differential expression analysis of RNA-seq data." BMC Bioinformatics **14**: 91.

Sonnemann, K. J., D. P. Fitzsimons, J. R. Patel, Y. Liu, M. F. Schneider, R. L. Moss and J. M. Ervasti (2006). "Cytoplasmic gamma-actin is not required for skeletal muscle development but its absence leads to a progressive myopathy." Dev Cell **11**(3): 387-397.

Sultan, M., M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert,

T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach and M. L. Yaspo (2008). "A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome." Science **321**(5891): 956-960.

Tarazona, S., F. Garcia-Alcalde, J. Dopazo, A. Ferrer and A. Conesa (2011). "Differential expression in RNA-seq: a matter of depth." Genome Res **21**(12): 2213-2223.

Tempfer, C. B., E. K. Riener, L. A. Hefler, J. C. Huber and A. Muendlein (2004). "DNA microarray-based analysis of single nucleotide polymorphisms may be useful for assessing the risks and benefits of hormone therapy." Fertil Steril **82**(1): 132-137.

van Wijk, E., E. Krieger, M. H. Kemperman, E. M. De Leenheer, P. L. Huygen, C. W. Cremers, F. P. Cremers and H. Kremer (2003). "A mutation in the gamma actin 1 (ACTG1) gene causes autosomal dominant hearing loss (DFNA20/26)." J Med Genet **40**(12): 879-884.

Walsh, T., H. Shahin, T. Elkan-Miller, M. K. Lee, A. M. Thornton, W. Roeb, A. Abu Rayyan, S. Loulus, K. B. Avraham, M. C. King and M. Kanaan (2010). "Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82." Am J Hum Genet **87**(1): 90-94.

Wang, J. L., X. Yang, K. Xia, Z. M. Hu, L. Weng, X. Jin, H. Jiang, P. Zhang, L. Shen, J. F. Guo, N. Li, Y. R. Li, L. F. Lei, J. Zhou, J. Du, Y. F. Zhou, Q. Pan, J. Wang, J. Wang, R. Q. Li and B. S. Tang (2010). "TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing." Brain **133**(Pt 12): 3510-3518.

Wang, K., M. Li and H. Hakonarson (2010). "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." Nucleic Acids Res **38**(16): e164.

Wang, L., Z. Feng, X. Wang, X. Wang and X. Zhang (2010). "DEGseq: an R package for identifying differentially expressed genes from RNA-seq data." Bioinformatics **26**(1): 136-138.

Yang, W., M. Ahmed, M. Elian, S. A. Hady el, T. S. Levchenko, R. R. Sawant, S. Signoretti, M. Collins, V. P. Torchilin and S. N. Goldberg (2010). "Do liposomal

apoptotic enhancers increase tumor coagulation and end-point survival in percutaneous radiofrequency ablation of tumors in a rat tumor model?" Radiology **257**(3): 685-696.

Zhu, M., T. Yang, S. Wei, A. T. DeWan, R. J. Morell, J. L. Elfenbein, R. A. Fisher, S. M. Leal, R. J. Smith and K. H. Friderici (2003). "Mutations in the gamma-actin gene (ACTG1) are associated with dominant progressive deafness (DFNA20/26)." Am J Hum Genet **73**(5): 1082-1091.

Zhu, M., T. Yang, S. Wei, A. T. DeWan, R. J. Morell, J. L. Elfenbein, R. A. Fisher, S. M. Leal, R. J. Smith and K. H. Friderici (2003). "Mutations in the gamma-actin gene (ACTG1) are associated with dominant progressive deafness (DFNA20/26)." American journal of human genetics **73**(5): 1082-1091.

국 문 초 록

새로운 기술의 발전은 그로부터 얻어진 자료를 분석하기 위한 새로운 방법론을 필요하게 한다. 대용량 자료의 시초라 할 수 있는 마이크로어레이 자료가 등장하고 이를 분석하기 위한 많은 방법들이 고안되어왔다. 유전체와 질병과의 상관관계를 분석하기 위한 전장유전체상관분석, 그룹간 차이를 보이는 유전자 발현 확인을 위한 분석 등, 다양한 생명현상의 원인을 발견하기 위해 여러 종류의 오믹스자료를 분석하는 방법들이 개발되어 온 것이다.

마이크로어레이 개발 초기에는 적은 수의 샘플에 대한 자료 분석방법이 개발되었으나 이후에는 많은 샘플수의 자료분석 방법론 개발이 주가 되어왔다. 저자는 그 이유를 마이크로어레이 기술의 다음 두 가지 특징으로부터 찾는다. 기술의 빠른 안정화와 기술의 불완전성. 안정화를 일찍 꾀한 마이크로어레이 기술은 빠르게 가격 경쟁력을 확보할 수 있었고 이는 많은 연구자들이 현실적으로 지불할 수 있는 많은 수의 자료를 생산할 수 있게 하였다. 두 번째는, 마이크로어레이 기술은 칩 위의 제한된 공간에 알고자 하는 정보를 '프로브(probe)'라는 서열로 미리 심어놓기 때문에 개별적 특징을 관찰하기 보다는 다수에서 관측될 수 있는 공통된 정보로부터 질병과 같은 관심 있는 현상을 설명해왔다.

차세대 염기서열해독(Next-generation sequencing, NGS)이라 불리는 기술은 태생적으로 개별적 정보에 관심을 갖는 기술이다. 많은 그룹의 노고로 진행된 서열해독을 통해 완성된 Human Genome Project의 시작부터 NGS와 '개인맞춤형'이란 단어는 궁합이 잘 맞는 조합이었다. 그러나 개별 시료가 가지고 있는 특

징을 1 염기서열의 해상도로 읽어내어 그것이 갖는 특징을 밝혀내는 작업은 결코 쉽지 않은 일이다. 뿐만 아니라 상대적으로 고비용의 NGS 기술은 연구자가 충분한 수의 자료를 생산하기 어렵게 하고, 결국 적은 수의 자료로부터 의미 있는 결과를 도출해 내는 것이 중요한 연구 주제가 되었다.

저자는 본 연구를 통해 서열해독을 통해 얻어진 소규모의 유전체와 전사체 자료를 통해, 자료가 가진 특징으로부터 얻을 수 있는 정보를 극대화할 수 있는 분석 방향과 방법을 개발하고자 한다. 본 연구와 동일한 목적으로 개발된 다른 방법들과의 차별성을 위해, 저자는 유전체 자료 분석을 위해서는 난청을 가진 가족의 전장 엑솜 유전체 자료를, 기존의 방법론을 이용하여 다양한 분석을 동시에 수행하여 원인 유전자를 찾는 소위, multiphasic 분석방향을 제시한다. 이러한 단계별 multiphasic 분석을 통해 적은 수의 가족데이터에서 멘델리안 유전병의 원인 유전자를 효율적으로 찾아낼 수 있음을 보이고자 한다.

또한 전사체 자료 분석의 경우, 저자는 반복수가 전혀 없는 두 조건하에서 얻어진 자료의 발현차 검정을 수행할 수 있는 방법을 제시하고 이를 여러 실제 RNA-Seq 자료와 모의실험 자료에 적용하여 반복 수가 적거나 혹은 전혀 없는 경우에도 저자가 제시한 방법이 의미 있게 발현차가 생긴 유전자를 찾을 수 있는 좋은 방법임을 보이고자 한다.

주요어: 차세대염기서열자료분석, 적은샘플자료분석, 유전자발현차검정,

학 번: 2006-30792